

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MÉTHODES BIO-INFORMATIQUES POUR L'ANALYSE DES MÉCANISMES
MOLÉCULAIRES CONDUISANT À LA RÉSISTANCE AUX MÉDICAMENTS
DANS LE CANCER DU SEIN

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR
ZAHIA AOUABED

JUILLET 2016

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.01-2006). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

J'exprime tout d'abord mes profonds remerciements à mon directeur de recherche Abdoulaye Baniré Diallo. Sa rigueur, sa justesse et ses riches enseignements m'ont été d'une grande utilité tout au long de mes travaux et de mes démarches méthodiques.

Je remercie également toute l'équipe du laboratoire bio-informatique de l'UQAM.

Je tiens aussi à remercier mon mari, ma sœur et mes parents. Votre soutien tout au long du projet a été très précieux.

TABLE DES MATIÈRES

REMERCIEMENTS.....	ii
LISTE DES FIGURES.....	vi
LISTE DES TABLEAUX.....	ix
LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES.....	x
RÉSUMÉ.....	xii
INTRODUCTION.....	1
CHAPITRE I	
PRINCIPES DE L'EXPRESSION DES GÈNES ET SA RÉGULATION.....	6
1.1. Biologie.....	6
1.1.1. Introduction à l'ADN, l'ARN et les gènes.....	6
1.1.2. Expression des gènes.....	9
1.1.3. La régulation génétique.....	11
1.1.4. Les facteurs de transcription et les régions régulatrices.....	12
1.1.5. Exemples de facteurs de transcription : Les récepteurs des oestrogènes.	15
1.1.5.1. Organisation structurelle et fonctionnelle du ER.....	15
1.1.5.2. Les différents domaines du ER.....	17
1.1.5.3. Les ligands du ER.....	18
1.1.5.4. Les voies de signalisation du recepteur ER.....	21
1.1.6. Analyse de la régulation génique.....	23
CHAPITRE II	
ANALYSE DE LA RÉGULATION GÉNÉTIQUE : APPROCHES	
ANTÉRIEURES.....	25
2.1. Le profil de l'expression de gènes pour l'identification des interactions de	
régulation.....	25
2.1.1. Identification du profil de l'expression de gènes grâce aux	
micropuces.....	25
2.1.2. Validation expérimentale des gènes cibles.....	28
2.1.3. Bio-informatique et régulation génétique.....	29

CHAPITRE III

ANALYSE DE LA RÉGULATION GÉNÉTIQUE : NOUVELLES

APPROCHES.....	41
PARTIE I : ChIP-Seq.....	43
3.1. Immuno-précipitation de la chromatine (ChIP).....	43
3.2. Identification des fragments issus de ChIP.....	44
3.2.1. Immuno-précipitation de la chromatine suivie de l'hybridation sur des micropuces (ChIP on chip).....	45
3.2.2. Immuno-précipitation de la chromatine suivie du séquençage haut débit (ChIP-seq).....	45
3.3. Définition du séquençage.....	46
3.4. Séquençage de la première génération.....	46
3.5. Séquençage haut débit (SHD).....	48
3.6. Destin des fragments séquencés : objectifs d'une expérience ChIP-seq.....	56
3.7. Chip-seq vs ChIP on chip.....	57
3.8. Considérations expérimentales dans une expérience ChIP-seq.....	59
PARTIE II : Analyse des données issues d'une expérience ChIP-Seq (workflow ChIP-Seq).....	64
3.9. Contrôle de la qualité (Qc).....	65
3.10. Alignement des séquences de reads.....	68
3.11. Identification de Pics "peak calling".....	87
3.12. Analyse des séquences.....	97
3.13. Les outils de visualisation.....	109
3.14. Les Mesures de qualité.....	110
3.15. Analyses en Aval.....	117
3.15.9.1. Problématique de l'analyse des liaisons différentielles.....	131
3.15.9.2. Approches actuelles pour l'analyse de la liaison différentielle dans ChIP-seq.....	136

CHAPITRE IV

NOUVELLE APPROCHE BIO-INFORMATIQUE POUR L'ANALYSE DES
LIAISONS DIFFÉRENTIELLES ASSOCIÉES À UN FACTEUR DE
TRANSCRIPTION.....

.....	139
4.1. Pipeline de la nouvelle approche.....	139
4.2. Application de la méthode aux données du cancer du sein.....	144
4.2.1. Contexte de l'étude.....	144
4.2.2. Objectif de l'étude.....	145
4.2.3. Problématique et motivation.....	146
4.2.4. Méthodologie.....	147
4.2.5. Matériels et Méthodes.....	148

CHAPITRE V	
RÉSULTATS ET DISCUSSION.....	189
CONCLUSION.....	237
BIBLIOGRAPHIE.....	263
GLOSSAIRE.....	241

LISTE DES FIGURES

Figure	Page
1 : Les différentes structures de l'ADN.....	9
2 : L'expression d'un gène eucaryote.....	10
3 : La régulation chez les eukaryotes.....	13
4 : Organisation modulaire des récepteurs nucléaires : exemple du récepteur des oestrogènes ER α	16
5 : Structure de l'estradiol (E2).....	18
6 : Les différentes voies de signalisation des oestrogènes, adaptée des travaux de (Heldring, 2007).....	21
7 : Analyse d'acides nucléiques par puce à ADN.....	26
8 : Description d'une expérience ChIP-Seq.....	44
9 : Description du séquençage par Synthèse.....	54
10 : Le workflow d'une expérience ChIP-Seq standard.....	64
11 : Schéma d'une stratégie d'alignement basée sur la table de hachage.....	72
12 : La transformation de Burrows-Wheeler pour les données des séquences génomiques.....	76
13 : Algorithmes de <i>MAQ</i> (Spaced seeds) et <i>Bowtie</i> (Burrows-Wheeler).....	78
14 : Profil des fragments d'ADN issus d'une expérience de ChIP séquencés à partir de leur extrémité 5'.....	89
15 : Schéma de la modélisation des pics pour une expérience Chip-Seq.....	91
16 : Programmes de "peak calling" sélectionnés pour évaluation.....	96
17 : La quantité de pics identifiés par différents peak callers.....	97
18 : PWM, construite à partir d'un alignement multiple local avec peu de gap, comme le modèle basique de TFBS.....	100

19 : Une image du navigateur de génome UCSC des enrichissements ChIP de NF- κ B et de H3K79me2 dans la lignée cellulaire lymphoblastoïde GM12878 de l'humain à partir des données ENCODE.....	110
20 : Une analyse d'échantillon de deux répliques pour le facteur de transcription NF- κ B d'une expérience ChIP-seq à partir de données ENCODE.....	113
21: Les scores PhastCons moyens au niveau des sites de liaison du TF utilisés pour évaluer la qualité des pics de ChIP-seq.....	114
22 : Visualisation de la distribution des pics H3K36me3 sur différentes catégories d'éléments.....	117
23 : La moyenne des signaux ChIP-seq de la H3K36me3 sur le metagène à différents niveaux d'expression des gènes dans la lignée cellulaire lymphoblastoïde humain GM12878.....	122
24 : Des exemples de pics différentiels putatifs après l'induction de la signalisation TLR4.....	134
25 : Choix des seuils sensibles et la comparaison des échantillons ChIP-seq.....	135
26 : Pipeline de la nouvelle approche proposée.....	148
27 : Exemple d'un fichier de type fastQ, contenant des données de séquençage	150
28 : Piste ucsc de l'expérience ER-LCC2-1_VS_ER-LCC9-2.....	173
29 : Description de la création d'un fichier de densité WIG (source UCSC).....	176
30 : Les quatre mesures de qualité (le nombre de lecture par position unique, l'auto-correlation des "tags" le biais de séquence et le biais GC) obtenues à l'aide du logiciel HOMER dans chacune de nos trois expériences ChIP-Seq.....	196
31 : Les diagrammes de Venn des ensembles de pics identifiés à partir des deux répliques individuelles au même cut-off pour le peak calling dans les cellules MCF7, LCC2 et LCC9.....	200
32 : Annotations des localisations des pics pour chacune des trois conditions....	204
33 : Distributions du nombre des pics à proximité des TSS (-10000, +10000	

bp) pour chacune des expériences (incluant les expériences replicats).....	208
34 : Le nombre de site de liaisons partagées entre nos trois conditions deux à deux.....	212
35 : Les résultats de l'analyse quantitative pour chaque paire de condition.....	217
36 : Les motifs consensus identifiés dans les sites de liaison ERE.....	221
37 : Extrait de la liste HTML des motifs enrichis (le top 12) obtenu avec le logiciel HOMER pour les pics dans la catégorie LCC2-1 vs MCF7-1 croissant ("increase").....	222

LISTE DES TABLEAUX

Tableau	Page
1 : Matrice de calcul.....	34
2 : Matrice de fréquence corrigée.....	35
3 : Matrice de poids (modèle de Bernoulli).....	37
4 : Contenu informationnel.....	39
5 : (a) Avantages et mécanismes des séquenceurs. (b) Composants et coût des séquenceurs. (c) Application des séquenceurs.....	50
6 : Evaluation de la qualité et traitement de reads.....	67
7 : Table des logiciels d'alignement.....	70
8 : Exemples de "peak callers" employés dans ChIP-seq.....	94
9 : Exemples d'outils pour l'analyse de motif des pics ChIP-seq et leurs utilisations.....	127
10 : Logiciels pour l'analyse de la liaison différentielle dans ChIP-seq	138
11 : Récapitulatif des principaux scripts décrivant les étapes d'analyse.....	188
12 : Statistiques de l'alignement des lectures et le nombre de pics ($FDR \leq 1\%$) identifiés.....	192
13 : Nombre de sites de liaison ERa identifiés dans les différentes conditions ($FDR \leq 1\%$).....	213
14 : Liste des motifs sur-représentés présents dans nos trois conditions avec le pourcentage des séquences cibles les contenant	223

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ADN : Acide DésoxyriboNucléique

AE : Anti-Estrogènes

ARN : Acide RiboNucléique

BWT : Transformation de Burrows-Wheeler

ChIP : Immuno-Précipitation de la Chromatine

CNV : Variation du Nombre de Copies

CRM : Modules de Régulation en Cis

EMSA: Essais de décalage de mobilité électrophorétique ("Electrophoretic Mobility Shift Assays")

ER : récepteur de l'estrogène ("Estrogen Receptor")

ERa : récepteur de l'estrogène alpha

ERb : récepteur de l'estrogène bêta

FDR : taux de fausse découverte ("False discovery rate")

FT : Facteur de Transcription

IC : contenu informationnel ("Information Content")

IDR : Irreproducible Discovery Rate

IP : Immuno-précipitation

NR : récepteurs nucléaires

NGS : Séquençage de nouvelle génération ("Next Generation Sequencing")

pb : Paires de bases

SNP : polymorphismes d'un seul nucléotidique

SHD : séquençage haut débit

TFBS: site de liaison du facteur de transcription ("Transcription Factor Binding Site")

TSS: site de début de transcription ("Transcription Start Site")

OS : système d'exploitation

BAM : fichier Sam binaire

Bustard "Illumina's base calling software": Logiciel d'appel de base de Illumina

CSFASTQ : le fichier FASTQ d'espace couleur (encodage SOLiD) ("color space (SOLiD encoded) FASTQ file"):

CSFASTA: le fichier FASTA d'espace couleur (encodage SOLiD) ("color space (SOLiD encoded) FASTA")

Eland : algorithme d'alignement Illumina

FASTA : format basé sur le texte pour la représentation des séquences nucléotidiques.

FASTQ : format basé sur le texte pour la représentation des séquences nucléotidiques et leurs scores de qualité.

QUAL FASTA : format basé sur le texte pour la représentation des scores de qualité.

SAM : Alignement des séquences /Map

SCARF : output du logiciel Gerald d'Illumina (fichier séparé par colonne avec un enregistrement par ligne contenant le nom du read, la séquence, et la qualité)

VCF : Format d'appel de Variant

BED : "Browser Extensible Data", un format de fichier basé sur le texte

DIVET : Format de fichier input séparé par des virgules de la variation Hunter

GFF : Format des caractéristiques générales

Maq : Format d'alignement propriétaire MAQ

QSEQ : Format de fichier de résultats d'appel de base d'Illumina

SOAP :Format d'alignement propriétaire SOAP

RÉSUMÉ

La technologie ChIP-Seq est utile pour l'étude de l'interaction protéine-ADN et ses données permettent d'analyser de façon efficace les mécanismes de la régulation génique, liés à un facteur de transcription à l'échelle du génome. Le traitement des quantités énormes de données générées par cette technologie nécessite des moyens informatiques puissants et efficaces. Mais, il n'existe en effet que peu ou pas de pipelines intégratifs offrant la possibilité de mener une analyse comparative de plusieurs expériences ChIP-Seq. Dans ce mémoire, nous abordons ce problème en concevant et implantant une plateforme d'analyse intégrée accessible à partir de Calcul Québec. Cette plateforme, exploitant en partie des paquets présents dans MUGQIC de Génome Québec, présente une approche originale et efficace pour traiter ce type de données. Elle a permis d'analyser les données ChIP-Seq du facteur de transcription ERa. Ce dernier est un facteur de transcription qui joue un rôle important dans le développement et la progression du cancer du sein. Trois lignées cellulaires de ce cancer ont été utilisées : une lignée sensible et deux lignées résistantes aux médicaments. Notre nouvelle approche a pu montrer que c'est le nombre des événements de liaison identifiés, leur localisation et l'intensité des pics au niveau des sites qui diffèrent entre les cellules sensibles et les cellules résistantes aux médicaments. Ces différences existent aussi entre les deux lignées résistantes aux médicaments. Avec l'analyse des motifs ainsi que les annotations fonctionnelles des sites identifiés, nous avons non seulement retrouvé le site de liaison du facteur de transcription d'intérêt ERa, mais nous avons également réussi à identifier des motifs rapportés dans la littérature comme se liant à la même région de liaison que ERa. L'analyse de l'enrichissement de GO a donné un aperçu de la fonction des gènes. Ces résultats ont indiqué que les gènes régulés par ERa sont liés au développement, la progression et les métastases du cancer du sein.

INTRODUCTION

Un organisme animal est caractérisé par un mode de fonctionnement très complexe. Ce dernier est sous le contrôle d'un réseau multicouche de mécanismes de régulation qui contrôlent étroitement le fonctionnement des cellules, des tissus et des organes. L'ensemble complet de gènes peut être considéré comme un réseau de régulation génique qui coordonne la prolifération et les processus de différenciation tout au long du développement d'un organisme. Ce réseau est sous le contrôle de petites molécules appelées facteurs de transcription (TFs), qui sont capables de réguler le niveau d'expression d'un certain nombre de gènes cibles en se liant à des sites spécifiques dans leurs régions promotrices. La plupart des génomes des vertébrés contiennent de 20.000 à 30.000 gènes qui codent pour des protéines dont environ 10% codent pour ces TFs.

Pour plus d'une décennie, l'identification des profils de l'expression des gènes en utilisant des puces (ChIP-chip) et plus récemment l'immuno-précipitation de la chromatine suivie du séquençage à haut débit (ChIP-Seq) ont été les principales méthodes pour analyser et élucider les réseaux de régulation de gènes. ChIP-Seq se distingue par le fait qu'elle permet une analyse *in vivo* des sites de liaison des TFs à l'échelle du génome. Comme les TFs se lient à l'ADN d'une manière spécifique à la séquence, une expérience ChIP-seq réussie doit être caractérisée par des sites présentant un enrichissement significatif du motif de liaison du TF d'intérêt. Les séquences motifs enrichies peuvent par la suite être identifiées avec des méthodes de découverte de novo ou en scannant les régions de sites aux motifs connus stockés dans des bases de données (Bailey, 2011 ; Ma, 2012; Zheng, 2015). Actuellement, plusieurs bases de données de motifs de liaison de TFs connus sont accessibles au

public, telles que JASPAR (Sandelin, 2004), UniPROBE (Robasky, 2011) ou TRANSFAC (Matys, 2003). En outre, avec l'apparition de nouvelles technologies et de données génomiques, des groupes expérimentaux et computationnels continuent à améliorer des motifs de FTs ou à caractériser d'autres restés jusqu'à lors inconnus (AlQuraishi, 2011 ; Nutiu, 2011; Zheng, 2015; Kulakovskiy, 2016).

Dans ce document, nous présentons une étude basée sur des données ChIP-Seq issues de trois lignées cellulaires du cancer du sein. Le cancer du sein constitue une étude de choix, car il est une des priorités de la communauté scientifique et médicale. Il s'agit d'une affection maligne multifactorielle dont les mécanismes sont, dans l'ensemble, encore mal compris. Cette incompréhension peut expliquer en partie le fait que la mortalité n'a pratiquement pas diminué depuis plus de 25 ans. En effet, malgré l'intérêt porté et les progrès dans les traitements prodigués, ce type de cancer demeure la seconde cause de mortalité par cancer au Canada (Comité directeur de la Société canadienne du cancer, 2010). Il constitue un grave problème de santé publique, puisqu'il touche 1 femme sur 10 et provoque 20 % des décès par cancer.

Les recherches antérieures ont démontré que les oestrogènes, qui sont des hormones naturelles, jouent un rôle important dans la genèse et la progression du cancer du sein. Ils exercent leur action via le récepteur des oestrogènes alpha (REa), un facteur de transcription capable de moduler la transcription de nombreux gènes impliqués dans la prolifération, la différenciation cellulaire ou l'apoptose. Afin de contrer les effets des oestrogènes et de ses récepteurs, de nombreuses molécules, appelées anti-oestrogènes (AE), tels que le tamoxifène et le fulvestrant, sont utilisées. Leur rôle est d'entrer en compétition avec l'oestrogène pour la liaison à ses récepteurs et de bloquer la prolifération des cellules cancéreuses mammaires. Malheureusement, les patientes traitées par ce type de molécules développent systématiquement une résistance aux thérapies anti-estrogène. Les facteurs de croissance qui entraînent une activation du ERa en l'absence d'oestrogène (activation de manière ligand indépendante) pourraient

être responsables de l'apparition de ces phénomènes de résistance. En effet, des études récentes ont montré une augmentation accrue de l'expression des facteurs de croissance dans les cellules résistantes aux AE (Osborne CK et al, 2011). De plus, le profil des sites de liaison des ERa à l'ADN, activés par ces facteurs de croissance, diffère de ceux induits par leur liaison à l'estrogène (Ross-Innes CS, 2012). Ainsi, l'analyse comparative de ces profils permettra de mieux comprendre les réseaux de régulation et les mécanismes moléculaires sous-jacents à l'acquisition de la résistance aux AEs dans le cancer du sein.

Dans ce mémoire, nous proposons une méthode bio-informatique efficace pour réaliser cette analyse. Elle se base sur des données de l'immuno-précipitation de la chromatine suivi du séquençage Illumina du facteur de transcription ERa entre trois lignées cellulaires du cancer du sein. Ces lignées correspondent à trois conditions de réponse aux médicaments : MCF7 sensible aux médicaments (tamoxifène et fulvestrant), la lignée LCC2 résistante au tamoxifène et la lignée LCC9 résistante au tamoxifène et au fulvestrant, à la fois. L'approche présentée repose sur une combinaison de trois étapes du pipeline MUGQIC de génome québec, spécialisé dans l'analyse d'une seule expérience ChIP-Seq, avec une partie d'analyse comparative intégrée et une phase préalable de traitement des données. Cette dernière traite le cas de séquences déjà alignées sur des versions plus anciennes du génome cible. L'analyse comparative, quant à elle, filtre et catégorise les sites de liaisons en deux étapes indépendantes, une méthode complète peu employée dans la littérature. La première étape caractérise le nombre de sites de liaison spécifiques à chaque condition ainsi que le nombre de sites de liaison partagés entre les conditions deux à deux. De son côté, la deuxième étape applique une analyse quantitative, basée sur le calcul de la densité des reads au niveau des sites de liaison, afin d'en extraire les liaisons différentielles. Une fois les sites des liaisons différentielles du facteur ERa

caractérisés, une annotation des localisations des sites par rapport à des régions annotés dans le génome (TSS, extrémité 5...) et une analyse de motifs ont été effectuées. Enfin, les annotations fonctionnelles des sites de liaison identifiés ont été réalisées en exploitant la puissance de l'ontologie de Gènes GO.

Ce document se compose de quatre chapitres principaux, une conclusion et une annexe :

Le chapitre I décrit les principes de base de l'expression et de la régulation de l'expression des gènes, suivi de la biologie du récepteur ERa comme exemple de facteur de transcription.

Le chapitre II donne un aperçu général des anciennes méthodes d'analyse de la régulation génique.

Le chapitre III présente l'expérience ChIP-seq comme nouvelle méthode d'analyse des réseaux de la régulation de l'expression des gènes. La description sera subdivisée en deux grandes parties. La première décrira l'expérience de l'immuno-précipitation de la chromatine (ChIP), les anciennes et les nouvelles méthodes du séquençage ainsi que les objectifs d'une expérience ChIP-Seq, ses avantages par rapport à une expérience ChIP-chip, ses faiblesses ainsi que ses considérations expérimentales. La deuxième partie, quant à elle, concernera la description du déroulement de l'analyse computationnelle des données issues d'une expérience ChIP-seq (*Workflow*). Elle décrira aussi la problématique posée lors de l'analyse des différences dans les réseaux de régulations entre plusieurs conditions (l'identification des liaisons différentielles entre plusieurs expériences ChIP-Seq) et donnera un aperçu des approches connues dans le domaine

Le chapitre IV présente notre nouvelle approche bio-informatique pour l'analyse des liaisons différentielles d'un facteur de transcription entre plusieurs types ou conditions cellulaires ainsi que son application sur des données réelles.

Le chapitre V fournit les principaux résultats obtenus suite à l'application de la méthode.

La section conclusion présente une synthèse basée sur les résultats obtenus par l'étude et les perspectives de recherche qui en découlent.

L'Annexe présente quelques résultats obtenus avec la deuxième répliquat utilisée tout au long de l'analyse (donne une évaluation de la reproductibilité des expériences à chaque étape de l'analyse), les tableaux liés aux annotations GO et KEEG, des sommaires des fichiers excel de sortie des peak calling ainsi que les matrices pour les motifs identifiés.

CHAPITRE I

PRINCIPES DE L'EXPRESSION DES GÈNES ET SA RÉGULATION

1.1. Biologie

Dans cette section nous présentons des notions de biologie moléculaire hautement vulgarisées qui sont en principe nécessaires pour bien mettre en contexte la technologie ChIP-seq. Les mécanismes et la nature des molécules impliquées sont beaucoup plus complexes. Une description plus complète est disponible dans différents manuels tels que celui de Harvey Lodish (Uzman, 2001).

1.1.1. Introduction à l'ADN, l'ARN et les gènes

Toutes les informations héréditaires d'un individu sont stockées dans le noyau de ses cellules sous forme d'acide désoxyribonucléique ou ADN. La découverte de cette désormais célèbre architecture en double hélice, support de l'information génétique, par James Watson et Francis Crick il y a plus de 60 ans avait marqué le début de la biologie moléculaire.

La macromolécule d'ADN est un polymère (une longue chaîne) formé par la concaténation de quatre nucléotides. Ces derniers sont des petites molécules simples composées d'un sucre, d'un phosphate et d'une base azotée qui les distingue en deux

catégories : d'un côté les purines avec l'adénine (A) et la guanine (G), de l'autre les pyrimidines avec la cytosine (C) et la thymine (T). C'est l'ordre de ces séquences nucléotidiques le long des chaînes qui détermine le message codé par l'ADN. De plus, grâce aux noyaux purines et pyrimidines, complémentaires par leurs propriétés physico-chimiques respectives (permettant leur appariement par des liaisons hydrogènes (liaisons faibles)), la formation d'un duplex de deux brins complémentaires, appelée *hybridation*, est énergétiquement favorable. Ce duplex a une tendance naturelle à s'enrouler sous la forme bien connue de double hélice (voir Figure 1). La caractéristique de complémentarité des bases est à l'origine du phénomène de réplication de l'ADN, à partir d'un polymère servant de patron, et expliquerait en partie l'origine de l'évolution à partir du premier ancêtre. Les structures primaires et secondaires de l'ADN sont, quant à elles, assurées par d'autres liaisons, dites les liaisons fortes. Celles-ci sont à l'origine de l'organisation des bases et de l'orientation 5'-3' des brins d'ADN. Les deux brins complémentaires ont de ce fait une orientation opposée.

L'ADN est présent dans le noyau de la cellule sous deux états : l'état déroulé comme durant la phase d'activité de la cellule (interphase) ou l'état condensé, par exemple, lors la division cellulaire. L'état déroulé correspond à la chromatine et se présente sous forme de filament d'ADN qui occupe la quasi-totalité du noyau. La chromatine est en réalité constituée de plusieurs filaments, elle serait d'ailleurs zonée pour éviter l'emmêlement entre filaments différents. De son côté, l'état condensé correspond au chromosome et se présente sous forme de filament d'ADN fortement enroulé sur lui-même (voir Figure 1). La condensation de l'ADN est rendue possible grâce à la liaison des filaments à de petites protéines, les histones. Ces protéines, associées par groupe de 4 pour former le nucléosome, sont réparties le long du filament et

présentent une structure semblable à celle d'un collier de perles, où, ces dernières sont agencées par l'enroulement du filament autour d'elles (voir Figure 1).

En interphase, le noyau de la cellule présente une chromatine plus ou moins condensée qui se présente sous forme de régions claires (l'euchromatine) et sombres (l'hétérochromatine). Il est important de noter que l'état condensé de l'hétérochromatine ne permet pas la transcription des gènes.

Contrairement à l'ADN, l'ARN est constitué d'une seule chaîne de nucléotides, plutôt que d'une chaîne double. En plus, dans sa composition, la thymine est remplacée par l'uracile (qui est également une pyrimidine). Les ARNs peuvent avoir trois grands types de fonctions : ils peuvent être des ARN messagers, support de l'information génétique d'un ou de plusieurs gènes codant pour des protéines (voir section expression des gènes), ils peuvent adopter une structure secondaire et tertiaire stables et accomplir des fonctions catalytiques (par exemple l'ARN ribosomique), comme ils peuvent servir de guide ou de matrice pour des fonctions catalytiques accomplies par des facteurs protéiques, comme c'est le cas par exemple des microARNs.

De son côté, un gène représente la zone de transcription d'ARN. Ainsi, chez les eucaryotes, elle correspondrait aux portions d'appariement entre ADN et ARN qui constituent les exons (les autres portions étant les introns). Cependant, sa définition peut être très complexe si l'on prend en compte l'ensemble des phénomènes biologiques identifiés au cours de ces dernières années [Gerstein *et al.*, 2007]. En effet, depuis sa première définition qui date de 1909 par De Vries, basée sur les concepts des généticiens Morgan et Mendel, le concept de gène n'a pas cessé d'évoluer. Toutefois, selon un critère de structure ou de fonction, une définition générale pourrait être élaborée. Ainsi, du point de vue structurel, portions d'ADN, les gènes correspondent à des régions d'un chromosome (voir Figure 1). Chaque gène a ainsi un emplacement précis sur le chromosome, identique quelque soit l'individu (d'une même espèce), que l'on appelle locus. Un gène n'est porté que par les

chromosomes d'une même paire. Chaque paire de chromosomes porte donc des gènes différents des autres paires. Du côté opérationnel, il constitue une unité responsable de la formation d'une molécule, et plus précisément d'un polypeptide (voir la section suivante: expression des gènes). Celui-ci participera à la formation de plus grosses molécules ou sera directement utilisable en tant qu'enzyme (qui interviendra dans la synthèse d'autres molécules), hormone, protéine d'ancrage.... Certains gènes, par l'intermédiaire du polypeptide qu'il code, ont une action régulatrice sur d'autres. Les cellules eucaryotes possèdent aussi beaucoup de gènes en double ou plusieurs exemplaires et certains totalement inutiles.

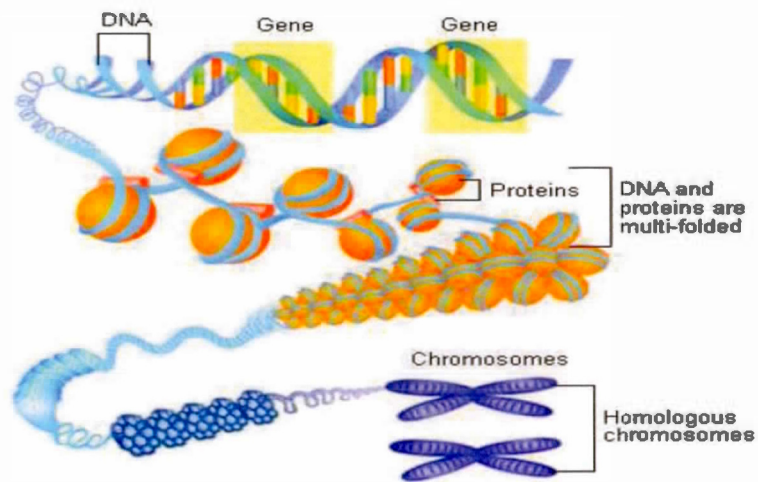


Figure 1: Les différentes structures de l'ADN (Alberts, 2002)

1.1.2. Expression des gènes

La cellule est l'unité de travail fondamentale de tout système vivant. Son fonctionnement est basé sur sa réaction à son environnement: elle réagit différemment en fonction des conditions et des stimuli. Cette interaction est rendue possible grâce à l'information contenue dans l'ensemble de ses gènes. Ainsi stockée, elle reste

disponible et se perpétue dans le temps à chaque fois qu'une cellule l'exploite. La réaction d'une cellule se traduit par l'expression (activation) ou la répression (extinction) des gènes. L'expression d'un gène permet à l'information qu'il contient d'être décodée en une molécule chimique appelée protéine. Cette dernière est responsable de la coordination des cellules et de la vie au niveau biochimique.

La protéine est synthétisée à partir de l'ADN suivant un processus à deux étapes: la transcription et la traduction schématisées dans la Figure 2. Le processus décrit concerne les gènes qui codent pour des protéines. En effet, certains d'entre eux codent pour des ARNr, ARNt ou pour d'autres petits ARN et suivent une autre voie de maturation qui ne rentre pas dans le cadre du projet.

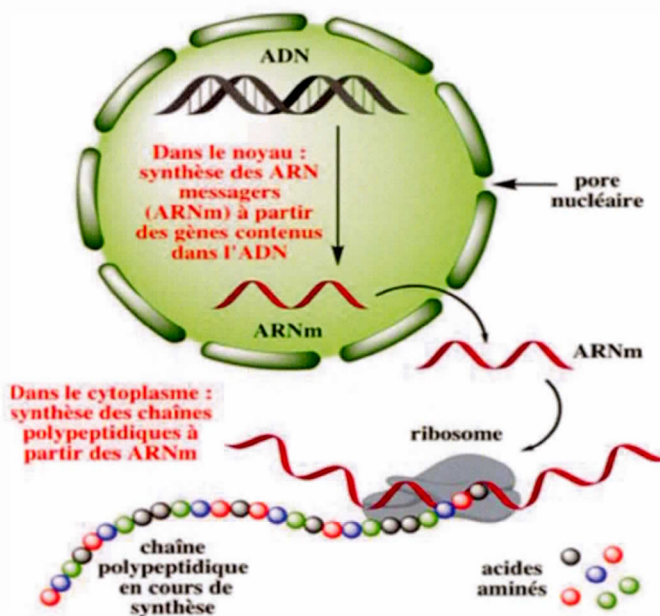


Figure 2: l'expression d'un gène eucaryote (E. Jaspard 2013).

La transcription est le processus par lequel l'information contenue dans la matrice d'ADN est copiée en ARN (Acide ribonucléique). Elle se passe dans le noyau de la cellule et est assurée par l'ARN polymérase. L'ARN messager (ARNm), ainsi

produit, subit plusieurs étapes de maturation avant d'être exporté dans le cytoplasme de la cellule. Dans le processus subséquent, l'étape de traduction en protéine s'enchaîne en décodant chaque triplet de nucléotides, nommé codon, porté par l'ARNm, en un acide aminé différent. Ce sont les différentes combinaisons de ces acides aminés qui définissent la nature de la protéine synthétisée.

Plusieurs mécanismes cellulaires assurent l'encadrement, la coordination et le bon fonctionnement de ces deux processus par la régulation de l'expression des gènes au bon moment et en fonction des conditions ou des circonstances physiologiques. Plus de détails concernant ce mécanisme sont présentés dans la section suivante.

1.1.3. La régulation génétique

Le patrimoine génétique d'un organisme particulier est a priori le même dans toutes les cellules qui le composent. Ce qui diffère d'un tissu à un autre, à travers le temps ou dans différentes conditions, est le programme de l'expression des gènes. Ceci signifie que différentes cellules dans différentes situations n'expriment pas les mêmes gènes, et si c'est le cas, il serait peu probable qu'ils soient exprimés dans les mêmes quantités.

Par exemple, dans une situation dangereuse et en réponse au stress (une condition particulière), les cellules des glandes médullosurrénales, situées au-dessus des reins, subissent une modification de l'expression des gènes en faveur de la transcription accrue de ceux impliqués dans la production de l'adrénaline. La présence d'adrénaline dans le sang déclenche instantanément des réactions dans tout le corps : le rythme cardiaque augmente, la respiration s'accélère, la pression artérielle s'élève, le cerveau et les muscles reçoivent plus d'oxygène, tandis que notre digestion ralentit et nos

pupilles se dilatent afin d'augmenter la vigilance. L'adrénaline permet de mobiliser l'organisme tout entier pour affronter le danger. Ainsi, l'expression des gènes est soumise à un processus de régulation génétique, une phase du contrôle de l'expression des gènes, qui comprend l'ensemble des mécanismes de régulation mis en œuvre pour permettre à la cellule de régler précisément la production de protéines en fonction de ses besoins.

Bien qu'elle puisse intervenir à différents niveaux de l'expression d'un gène : transcriptionnel, post-transcriptionnel, traductionnel et post-traductionnel, la régulation transcriptionnelle est le processus principal, de ce fait le plus étudié et qui sera concerné par le projet. Ce type de régulation dépend essentiellement de petites molécules appelées facteurs de transcription (FT).

1.1.4. Les facteurs de transcription et les régions régulatrices

Un facteur de transcription est une protéine qui a une affinité particulière pour certaines courtes séquences d'ADN, sur les quelles, elle peut se fixer grâce à sa structure 3D spécifique. Chez les eucaryotes, ces séquences, dites facteurs cis régulatrices, se situent à proximité du site d'initiation de la transcription (TSS pour « Transcription Start Site ») formant le promoteur proximal.

Le promoteur ne constitue pas la seule région régulatrice (Geyer, 1992) (voir figure 3). Il existe d'autres types comme:

- les activateurs "enhancers" qui sont capables de potentialiser l'action du promoteur à distance du site d'initiation de la transcription (TSS),
- les silencers distants qui ressemblent aux enhancers mais qui jouent le rôle de répresseurs, et

- les insulateurs qui correspondent à une séquence régulatrice affectant l'interaction entre l'enhancer et le promoteur.

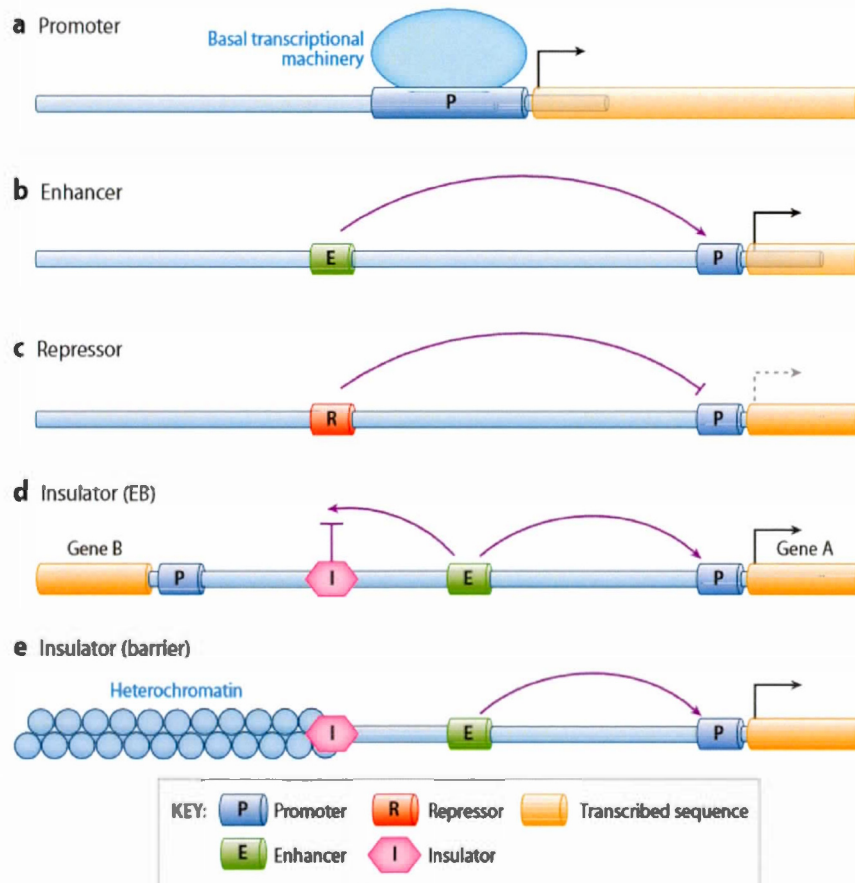


Figure 3 : La régulation chez les eukaryotes.

Schéma des Classes d'éléments régulateurs connus permettant la régulation de l'expression des gènes (Noonan, 2010) : (a) la séquence du promoteur montre la liaison de la machinerie générale de la transcription pour médier le contrôle de la transcription basale d'une séquence transcrite. (b) les séquences Enhancer et (c) represser montrent la médiation des effets positifs et négatifs de la transcription via l'interaction avec la séquence du promoteur. (d) les insulateurs de blocage des éléments Enhancer "Insulator (EB)" bloquent l'activation interdite du gène B par les éléments

régulateurs du gène A (enhancer blocking, d), et (e) les insulateurs barrière "insulator (Barrier)" empêchent la propagation de la condensation de la chromatine d'envahir les séquences régulatrices qui contrôlent un gène hypothétique (barrier, e).

Ces courtes séquences régulatrices de 6 à 15 paires de bases, connues sous le nom d'éléments régulateurs (RE) ou sites de fixation de facteurs de transcription (TFBS, «Transcription Factor Binding Site »), sont particulièrement conservées au cours de l'évolution.

Après la fixation aux TFBS, certains domaines du facteur de transcription mènent à une activation ou à une inhibition de la transcription d'un gène cible, jouant ainsi sur son taux de transcription. Cette modulation se fait en facilitant ou en rendant difficile la fixation du complexe d'initiation de la transcription (la modulation de l'accessibilité à la molécule d'ADN par une acétylation ou par une désacétylation des histones), ou en augmentant ou en diminuant le taux de recrutement de la polymérase (la modification de la stabilité de la liaison entre l'ADN et l'ARN polymérase).

Notez qu'approximativement 10 % de l'ensemble des 20 000 à 30 000 gènes existants codent pour des facteurs de transcription et qu'un facteur de transcription peut activer sa propre transcription dans certaines conditions et la réprimer dans d'autres (Shefer-Vaida, 1995). De plus, les facteurs de transcription n'agissent pas de manière indépendante, mais forment des complexes avec d'autres facteurs de transcription et cofacteurs protéiques (Fedorova, 2010). Ensemble, ils se lient au niveau des éléments régulateurs (RE) et sont communément appelés des modules de régulation en cis (CRM). Afin que la transcription du gène ait lieu, il est nécessaire que chacun des facteurs de transcription se fixe à la molécule d'ADN (Carroll, 2006). Parfois, la présence de cofacteurs ou de certaines protéines, comme éléments intermédiaires, est indispensable à l'initiation de la transcription. Accueillir ces intermédiaires exige cependant une organisation minutieuse et une conformation adéquate de chacun des facteurs.

En résumé, la régulation de la transcription est un mécanisme complexe et sa modulation nécessite l'action coordonnée de plusieurs facteurs de transcription (CH, 1998).

1.1.5. Exemples de facteurs de transcription: Les récepteurs des oestrogènes

Les récepteurs des oestrogènes (ER alpha et ER bêta) sont des facteurs de transcription dont l'influence sur la régulation des gènes dépend de la liaison de leurs ligands (voir la section ligands du ER). Ce sont des récepteurs stéroïdiens appartenant à la superfamille des récepteurs nucléaires (NR) (les autres membres étant les récepteurs des glucocorticoïdes, des androgènes, de la progestérone, de l'acide rétinoïque, des minéralocorticoïdes, et de la vitamine D).

Le récepteur ER alpha a été le premier à être découvert dans les années 1960. Il fut cloné en 1985 (Walter, 1985), séquencé en 1986 puis utilisé comme un marqueur pour le diagnostic et le traitement du cancer du sein (Green, 1986; Osborne, 1998). La découverte du ER bêta, viendra un peu plus tard, en 1995, et permettra d'éclaircir certains effets des oestrogènes (Mosselman, 1996). Cependant, son implication dans le développement du cancer du sein reste moins caractérisée.

1.1.5.1. Organisation structurelle et fonctionnelle du ER

Les récepteurs ER alpha et ER bêta sont codés par des gènes localisés sur des chromosomes différents, le 6 et le 14, respectivement (Mosselman, 1996; Menasce,

1993). Ces récepteurs présentent une organisation modulaire du domaine caractéristique des RNs et partagent une haute homologie du domaine protéique. Ils comprennent 6 régions fonctionnelles distinctes (A-F) (Herynk, 2004; Hewitt, 2002; Kumar, 1987) avec différents degrés de conservation entre elles. Le pourcentage d'identité des séquences en acides aminés entre le récepteur ER alpha et ER bêta est représenté dans la Figure 4.

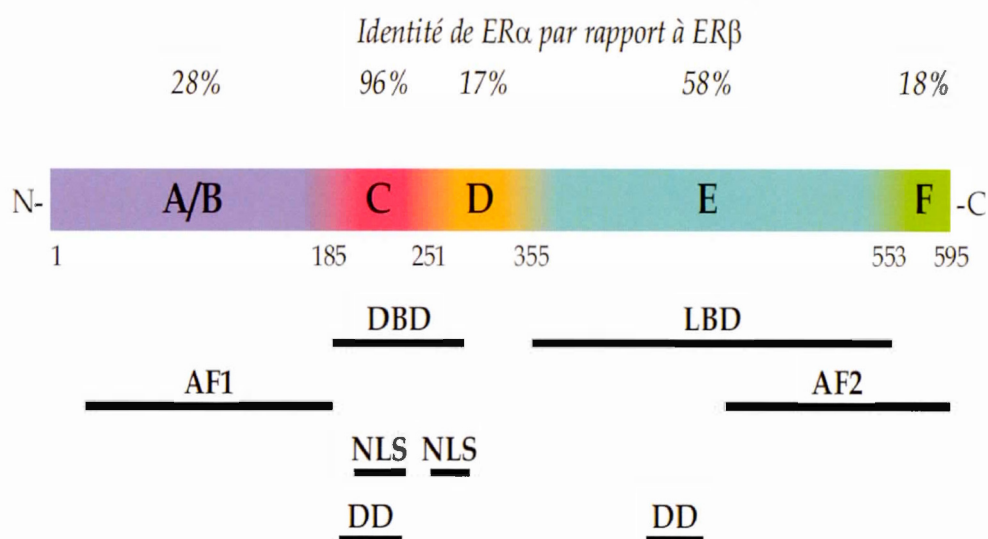


Figure 4: Organisation modulaire des récepteurs nucléaires: exemple du récepteur des oestrogènes ER α . Les récepteurs nucléaires présentent une organisation modulaire en lien avec leur fonction : le DBD et les fonctions de transactivations AF leur permettent de lier l'ADN et d'agir en facteur de transcription, tandis que le LBD place leur fonction sous la dépendance de leurs ligands respectifs. ER alpha et ER bêta partagent 96 % d'identité d'acide aminé dans le DBD, approximativement 53 % d'identité d'acide aminé dans le LBD et 30 % ou moins dans d'autres domaines, impliqués dans la transactivation et la localisation AF1 : fonction de transactivation 1, AF2 : fonction de transactivation 2, DBD : domaine de liaison à l'ADN, LBD : domaine de liaison au ligand. (Krust, 1986).

1.1.5.2. Les différents domaines du ER

Le premier est le domaine A/B N-terminal qui renferme des sites pour la phosphorylation et la fonction d'activation de la transcription ligand indépendante (AF-1) (participant à l'activation de la transcription du ER) (Gronemeyer, 1991; Tora, 1989). Cette région et la région N-terminale ne représentent que 28% et 18 % d'homologie entre ER alpha et ER bêta, respectivement, formant ainsi la zone la moins conservée. Ce sont ces différences au niveau de la région N-terminale qui seraient à l'origine des réponses spécifiques à chacun des récepteurs en fonction du type cellulaire, du promoteur, ou du ligand (McInerney, 1998).

Le second est le domaine C. Il comprend le domaine de liaison à l'ADN (DBD). Il représente la région la plus conservée avec 96% d'homologie entre ER alpha et ER bêta. Il est composé de deux doigts de zinc à quatre cystéines et permet aux récepteurs de se lier au niveau de séquences spécifiques de l'ADN, appelées les éléments de réponse aux oestrogènes (EREs) (Mader, 1993).

Le troisième domaine, noté D, est une région charnière non structurée et non conservée (Gronemeyer, 1995). Elle est constituée de signaux de localisation nucléaire hormonoindépendants (Ylikomi, 1992).

La région E, quant à elle, contient le domaine de liaison au ligand (LBD) et la fonction d'activation de la transcription dépendante du ligand (AF-2) (Tora, 1989). C'est une région bien conservée et impliquée dans la localisation nucléaire, la dimérisation du récepteur (Kumar, 1988; Tamrazi, 2002) et l'interaction avec les coactivateurs transcriptionnels (Henttu, 1997; White, 1997).

Enfin, le domaine F, peu conservé, est impliqué dans le recrutement de corégulateurs et probablement dans les mécanismes d'action des antagoniste (Montano, 1995; Nichols, 1998; Schwartz, 2002).

1.1.5.3. Les ligands de ER

1.1.5.3.1. Les oestrogènes

L'estrogène est le principal ligand naturel du récepteur des oestrogènes (voir figure 5). C'est une hormone stéroïdienne lipophile dérivée du cholestérol et sécrétée par les ovaires durant la période reproductive de la femme. Il traverse par diffusion la membrane cellulaire et nucléaire et se lie avec une haute affinité et spécificité à ses récepteurs.

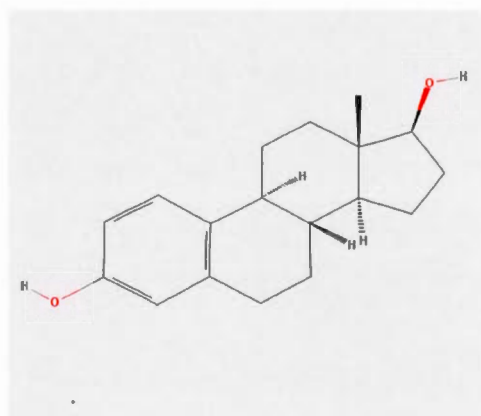


Figure 5: Structure de l'estradiol (E2) (<http://pubchem.ncbi.nlm.nih.gov>)

Les estrogènes jouent un rôle essentiel dans divers tissus cibles, y compris l'appareil reproducteur et le système nerveux central, vasculaire, et squelettique (le maintien de la densité et de la minéralisation osseuse) ainsi que dans le développement des glandes mammaires normales. Malheureusement, ils sont aussi fortement impliqués dans le développement et la progression du cancer du sein.

– Les estrogènes, les ERs et le développement du cancer du sein

À l'origine, les oestrogènes et ses récepteurs (ERs) sont des régulateurs essentiels de la prolifération des cellules épithéliales du sein, de la différenciation et de l'apoptose. Les ER alpha et ER bêta montrent des distributions tissulaires et des fonctions distinctes. Dans le sein normal, seulement 10 à 20 % des cellules épithéliales expriment ER alpha mais des délétions ciblées d'un ou des deux gènes ER, chez les souris, ont établi que ER alpha est le régulateur clé du développement de la glande mammaire, et il est responsable de bon nombre des effets des oestrogènes sur les tissus mammaires normaux et malins. L'exposition à vie au 17 bêta-estradiol (E2) ou "composés oestrogéniques" constitue un facteur de risque majeur pour le développement du cancer du sein. En fait, dans le tissu mammaire malin, l'action de ces oestrogènes se trouve dérégulée, ce qui entraîne un changement de la prolifération sans différenciation ou apoptose. Le pourcentage des cellules épithéliales exprimant ER alpha est d'ailleurs significativement augmenté sous ces conditions et la concentration locale en oestrogènes est dix fois plus élevée dans les tissus cancéreux que dans les tissus normaux (Navakauskiene, 2004).

1.1.5.3.2. Les anti-oestrogènes

Les anti-oestrogènes (ou antagonistes des oestrogènes) ont pour but de contrer les effets cancérigènes des oestrogènes. Ils se lient aux ERs de façon similaire à ces derniers, mais induisent une conformation différente du domaine de liaison au ligand (Leclercq, 2006; Wiseman, 2001). Ceci mène à la perte du recrutement des coactivateurs par le domaine AF-2 et le blocage de l'accès d'autres ligands.

Le plus grand défi dans le domaine du développement du médicament est de trouver des molécules capables de cibler la tumeur, en jouant le rôle antagoniste, tout en préservant la fonction agoniste bénéfique dans les autres tissus. Ainsi, plusieurs antagonistes ont été développés. Certains sont dits agonistes partiels du ER ou des SERMs pour "Selective Estrogen Receptor Modulators". Ces molécules possèdent une activité agoniste partielle dépendante du type cellulaire. Les SERMs les plus communs dans le traitement du cancer du sein sont le tamoxifène et le raloxifène. Le tamoxifène a un effet agoniste bénéfique sur les os chez les femmes post-ménopausées (Freedman, 2001; Kinsinger, 2002). Toutefois, son action agoniste dans d'autres tissus peut être néfaste comme dans l'utérus où il peut augmenter le risque du cancer de l'endomètre. D'autres antagonistes purs, appelés "SERDs" pour selective estrogen receptor down-regulator, ne possèdent aucune action agoniste (Wakeling, 1991). Dans ce cas, l'administration de traitement contre l'ostéoporose s'avère indispensable pour préserver la densité osseuse des patientes. Le plus commun des "SERD" est le Fulvestrant / ICI 182 780. Il exerce son action en se liant et en bloquant l'activité ER par son exportation en dehors du noyau et sa dégradation (Osborne, 2004).

1.1.5.3.3. Les phyto-oestrogènes

Les phyto-oestrogènes sont des composés dérivés des plantes avec les mêmes propriétés structurales que les oestrogènes et ont une action antagoniste partielle au niveau des tissus cibles.

1.1.5.4. Les voies de signalisation du recepteur ER

Les effets du récepteur ER résultent de son activation. Cette dernière peut être dépendante ou indépendante du ligand (estrogène). Une fois activé, ER peut influencer la transcription de gènes par plusieurs voies de signalisation (voir figure 6).

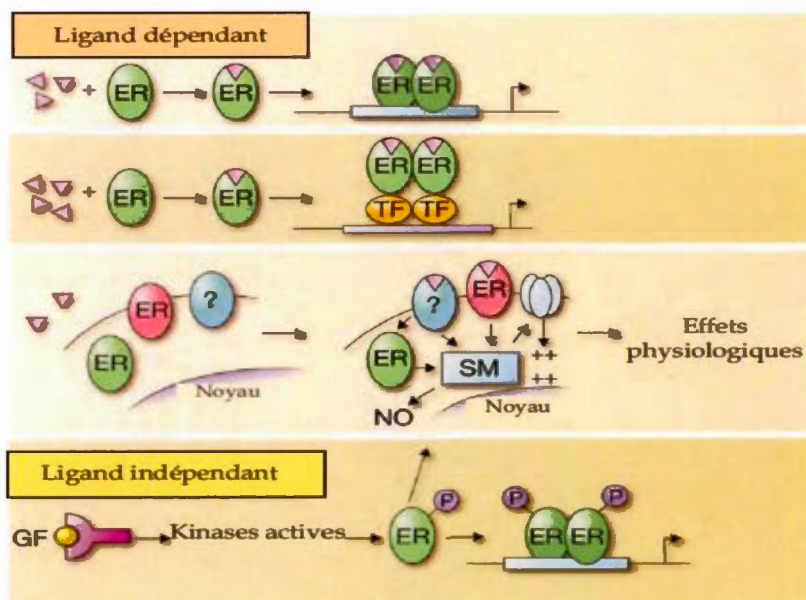


Figure 6: Les différentes voies de signalisation des oestrogènes, adaptée des travaux de (Heldring , 2007).

– Activation dépendante du ligand

- La voie classique est une voie génomique dépendante des éléments de réponse aux estrogènes (EREs). Elle implique l'activation de ER par liaison du ligand et mène à la régulation des gènes cibles par fixation directe sur des séquences ADN régulatrices spécifiques (les EREs).
- Les ERs activés par leur ligand peuvent aussi moduler la transcription de façon indirecte. Dans ce cas, ils se lient à d'autres facteurs de transcription par des interactions protéines-protéines, par exemple aux familles de facteur de transcription AP-1 et Sp-1 (Schultz, 2005; Safe, 2005). Ces derniers se lient au niveau de leur site de fixation à l'ADN et créent ainsi une action indirecte entre le récepteur ER et l'ADN (voir figure 6).
- La voie non-génomique est responsable des effets rapides des oestrogènes. L'activation des ERs, localisés au niveau des membranes cellulaires, mène à l'activation des voies de transduction du signal dans le cytoplasme (Chung, 2002; Kahlert, 2000). Cette cascade de transduction induit une réponse physiologique, produite par ER ou par un second messenger, sans pour autant impliquer la transcription de gènes (figure 6) (Heldring, 2007).

– Activation indépendante du ligand

- La voie de signalisation non-génomique à génomique mène à la régulation des gènes cibles en l'absence de ligand (ligand-indépendante). Elle active les récepteurs ERs par phosphorylation en réponse aux signaux des voies de signalisation des facteurs de

croissance et via l'activation des protéines kinases, telles que les kinases HER2-regulated mitogen-activated protein (MAP), ERK1 et ERK2 (Bunone, 1996). Après phosphorylation, le ER sera activé menant à sa dimérisation, sa liaison à l'ADN et enfin la régulation de la transcription (figure 6).

1.1.6. Analyse de la régulation génique

Afin d'établir le réseau de régulation des gènes, à savoir les gènes sur-régulés vs les gènes sous-régulés en réponse à un facteur de transcription, dans un type cellulaire donné, l'établissement du cistrome et/ou du transcriptome pour le profil de l'expression de gènes ont été les principales approches depuis de nombreuses années.

– Le cistrome

Un cistrome est un terme utilisé pour définir, in vivo et à l'échelle du génome, un ensemble de sites de liaison à l'ADN (éléments cis-régulateurs) recrutant un certain facteur de transcription (facteurs-trans).

Par exemple, le cistrome ER comprend l'ensemble des sites de liaison ER associés au modèle classique d'action de ER, ainsi que ceux associés à un autre facteur de transcription comme intermédiaire (liaison indirecte à l'ADN) (voir, ci-dessus, la section voies de signalisation du facteur de transcription ER).

- Le transcriptome

Le transcriptome définit l'ensemble des molécules ARNs, incluant l'ARNm, l'ARNr, l'ARNt, et les ARNs additionnels non-codants présents dans une cellule ou une population de cellules à un moment donné.

CHAPITRE II

ANALYSE DE LA RÉGULATION GÉNÉTIQUE : APPROCHES ANTÉRIEURES

Dans ce chapitre, nous décrivons les anciennes approches utilisées pour établir et analyser le réseau de régulation des gènes en réponse à un facteur de transcription donné.

Pour commencer, dans la section suivante, une brève description du principe de l'identification du profil de l'expression génique est donnée. Les faiblesses de la méthode ainsi que les raisons de la nécessité du cistrome y sont exposées.

2.1. Le profil de l'expression de gènes pour l'identification des interactions de régulation

2.1.1. Identification du profil de l'expression de gènes grâce aux micropuces

Afin d'identifier le réseau de la régulation des gènes, l'identification du profil de l'expression génique grâce aux microarrays a été la technique de choix depuis plusieurs années. De façon générale, après avoir effectué une modification d'un des composés de l'expérience (tel l'ajout ou la suppression d'une hormone, d'un médicament, ...), la technique se base sur la comparaison des mesures de l'expression des gènes entre deux expériences dont l'une est un contrôle. Ainsi, dans le cas de l'étude des gènes régulés par un FT, les changements trouvés dans l'expression de certains gènes, en réponse à des variations dans l'activité du facteur en question,

seront retenus et les gènes associés seront les cibles potentielles de ce dernier (voir Figure 7).

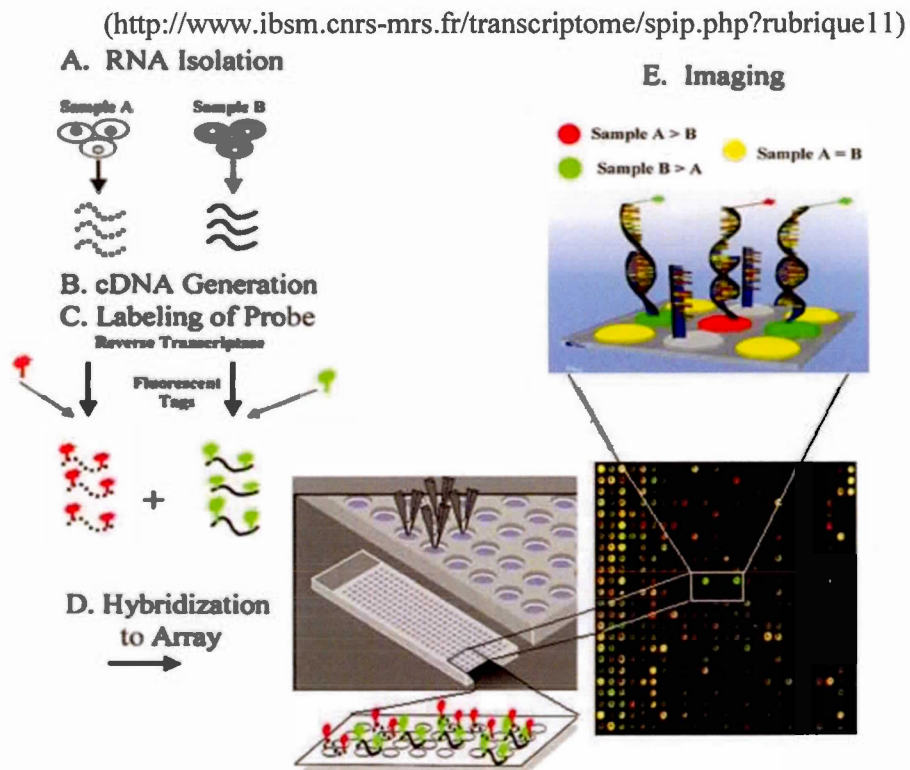


Figure 7 : Analyse d'acides nucléiques par puce à ADN.

Malgré l'engouement autour de cette technique et les nombreuses publications basées sur ses résultats, le profil de l'expression des gènes reste limité. En effet, il ne peut malheureusement pas discriminer une sur-régulation d'un gène cible putatif résultante d'une régulation directe et celle issue d'une régulation indirecte. Plus encore, le facteur de transcription pourrait agir comme un répresseur d'un gène intermédiaire agissant sur la transcription et le résultat est une sur-régulation conséquente du gène présumé comme cible.

Ainsi, la complexité des scénarios qui peuvent se produire dans le réseau de régulation est telle que l'établissement du profil de l'expression des gènes est insuffisant pour élucider le lien entre un facteur de transcription et ses gènes cibles. Ce lien, traduisant une régulation directe, est basé entre autre sur la liaison du facteur de transcription dans la région promotrice proximale du gène cible candidat. Même si d'autres régions distantes, les enhancers et les silencers, ont un rôle bien établi dans les réseaux de régulation, l'étude de la région promotrice proximale permet à elle seule d'élucider le mécanisme de régulation sous-jacent. Ceci est vérifié dans les études des promoteurs seuls (Suzuki, 2003; Guo, 2008), mais aussi dans des analyses à l'échelle du génome où une corrélation apparaît entre les profils d'expression de gènes et des motifs au niveau de leurs régions promotrices proximales (Roeder, 2009). Ainsi, la détermination du «cistrome» est devenue nécessaire pour caractériser les gènes qui sont directement régulés par un facteur de transcription (Zhang, 2008).

Deux approches pour l'étude des associations globales entre les facteurs de transcription et leurs gènes cibles prédits existent. La première est expérimentale pour la validation et la seconde est basée sur des modèles computationnels pour les liaisons des facteurs de transcription. Dans ce qui suit une présentation des deux stratégies.

2.1.2. La validation expérimentale des gènes cibles

Habituellement, l'approche expérimentale utilisée pour valider les résultats du profil de l'expression des gènes est le dosage d'une protéine appelée la luciférase. Le gène de cette protéine peut être isolé et intégré dans un petit morceau d'ADN, un plasmide, qui sera par la suite transféré dans une ou plusieurs cellules.

Il existe différents moyens d'insérer un plasmide, par exemple par des lipides, des sels, ou la force électrique. Le but étant de faire des trous à travers lesquels la luciférase plasmide pourrait entrer dans la cellule et produire la luciférase. Cette technique de la biologie moléculaire, connu sous le nom de la transfection, permet de doser la luciférase pour caractériser et évaluer la force de différents promoteurs dans les plasmides luciférase (car c'est le promoteur qui contrôle la quantité de luciférase produite par le gène). Ainsi, la force d'un promoteur est déterminée par le taux de bioluminescence produite par la transfection.

Les dosages de la luciférase trouvent une application dans la validation des cibles candidates retrouvées suite au profilage de l'expression des gènes après le knockouts des facteurs de transcription (Hecht, 2007; Young, 2007). Ils permettent de confirmer l'expression différentielle ainsi que de trouver (mapping) la position du contact protéine-ADN dans la région promotrice (Suzuki, 2003; Guo, 2008; Yamauchi, 2008).

Dans une première étape, des candidats de taille différente, construits et situés à des endroits distincts de la région promotrice présomptive, sont clonés à l'avant d'un gène rapporteur de la luciférase. Ils comportent une séquence promotrice minimale et seront transfectés dans une lignée cellulaire avec un second vecteur d'expression portant le FT régulateur candidat. Ainsi un signal peut être détecté, sous forme de

luminescence, si et seulement si, la séquence promotrice clonée a une activité de régulation. À la fin, seul le "construct" produisant la plus forte activité de la luciférase est considéré pour la prochaine analyse. La seconde étape vise à cibler le site réel de liaison du facteur de transcription. En effet, la taille de la région obtenue à l'issue de la première étape peut être grande (de l'ordre de 100 pb) et une recherche par des tests de matching sur des motifs consensus déjà connus s'avère nécessaire. Malheureusement, cette approche n'est pas toujours possible: dans certains cas, le facteur se lie de façon indirecte à l'ADN, en interagissant avec un autre facteur de transcription (voir section 1.1.5.4. du chapitre I), qui à son tour, se lie peut-être à un motif totalement différent (Suzuki, 2003). En plus, des motifs consensus connus n'est pas une donnée toujours disponible, par exemple, lorsqu'on procède à une recherche de novo.

2.1.3. Bio-informatique et régulation génétique

2.1.3.1. Approches computationnelles pour l'identification des motifs de liaison facteurs de transcription-ADN

Si on présume que le groupe de gènes co-régulés identifiés dans le profil d'expression génique dépend des mêmes facteurs de transcription, alors les régions promotrices en amont de ces gènes contiendraient automatiquement les sites de liaison correspondants à ces facteurs. À partir de ce groupe de gènes, des méthodes bio-informatiques pour prédire ces régions de courtes séquences surreprésentées ont été développées. Toutefois, avant d'arriver à cette étape d'analyse computationnelle, ces domaines de liaison des facteurs de transcription doivent être déterminés par des méthodes expérimentales.

Méthodes expérimentales

Un facteur de transcription peut avoir un ensemble de séquences ADN similaires qu'il peut reconnaître, dites motifs de reconnaissance. La détermination de ces motifs peut se faire par quelques méthodes expérimentales utilisées depuis des décennies dans l'étude de la régulation transcriptionnelle. Ces méthodes peuvent être classiques ou à haut débit. Dans ce qui suit une description brève des plus communément utilisées.

– Méthodes classiques

Les méthodes classiques, contrairement aux méthodes à haut débit (voir plus bas), ne peuvent analyser qu'un seul gène ou protéine à la fois.

- Retard sur gel

Le retard sur gel (electrophoretic mobility shift assays, EMSA) est l'une des techniques les plus importantes pour l'étude de la régulation des gènes et la détermination des interactions Protéine-ADN. Elle date du début des années 80 (Fried, 1981; Garner, 1981) et est basée sur l'observation suivante : lorsqu'ils sont soumis à une électrophorèse sur gel d'agarose ou de polyacrylamide, et en conditions non dénaturantes, les complexes Protéine-ADN migrent plus lentement que les fragments d'ADN linéaires seuls (non liés) (Hendrickson, W. (1985). *BioTechniques* 3, 346-354, - Revzin, A. (1989). *BioTechniques* 7, 346-354). Comme la fixation des facteurs de transcription ralentit la migration du fragment d'ADN, la détermination de leurs sites de liaison repose sur la simple comparaison entre la migration du complexe Protéine-ADN et l'ADN seul (non-complexé).

La technique de retard sur gel n'est pas limitée aux interactions protéine-ADN. En effet, en plus de permettre l'étude de nombreuses propriétés de ce type d'interactions (Gerstle, 1993) (par exemple mesurer l'affinité d'une protéine pour une séquence d'ADN ou déterminer le domaine de fixation de l'ADN sur une protéine), les interactions protéine-ARN (Allen, 1999) et protéine-peptide (Grunwald, 1998) peuvent aussi être étudiées en respectant le même principe de base.

- Empreinte à la Dnase I

L'expérience empreinte à la Dnase I fait généralement suite à la technique de retard sur gel. Elle permet de déterminer, avec une haute précision, les bases de la séquence promotrice qui interagissent physiquement avec le facteur de transcription.

La Dnase I est en fait une endonucléase capable de cliver, de manière non spécifique, les liens phosphodiester de l'un des deux brins d'ADN. La présence d'un complexe protéique (le facteur de transcription dans notre cas) sur la séquence d'ADN protège ce dernier par encombrement, et empêche l'endonucléase I de couper au niveau de sa séquence de fixation. Ainsi, en comparant les bandes résultantes de l'activité de l'endonucléase I sur un fragment d'ADN purifié et un fragment d'ADN lié à un facteur de transcription, on pourrait identifier une région sans bandes correspondant au site de fixation du facteur en question (Galas, 1978).

L'empreinte à la DNase I est très précise car elle peut identifier des interactions ADN-protéine qui peuvent passer inaperçues avec d'autres méthodes (Leblanc, 2001). Son plus grand inconvénient est qu'elle nécessite une quantité plus importante de protéines.

– Méthodes à haut débit

Avec l'apparition des premiers séquenceurs (voir section 3.4 du chapitre III), la biologie moléculaire a connu un développement spectaculaire. De nouvelles versions haut débit des méthodes expérimentales ont vu le jour. L'intégration de l'automatisation et de l'informatique a accéléré le travail en recherche en facilitant le stockage, le traitement, l'analyse et l'interprétation des données.

- SELEX

Le SELEX (pour Systematic Evolution of Ligands by EXponential enrichment) explore et identifie les séquences de fixation d'une protéine à l'ADN. La procédure se base sur un pool aléatoire de séquences produit par PCR avec des oligonucléotides synthétisés. La librairie de séquences peut aussi provenir du fractionnement d'un génome (SELEX génomique). Dans ce cas, les positions de liaison des protéines d'intérêt sont identifiées directement sur leurs séquences génomiques. SELEX se déroule en cycles répétés de sélection de séquences fixant la protéine (par exemple par la technique de retard sur gel (EMSA)), et d'amplification par PCR. La répétition de ce cycle permet l'enrichissement de la librairie d'oligonucléotides en séquences de haute affinité pour la protéine d'intérêt (Tuerk, 1990).

Les méthodes expérimentales sus-citées produisent un nombre important de séquences liées qui nécessitent des approches computationnelles pour les analyser.

2.1.3.2. Analyse des séquences liées (ou séquences cis-régulatrices)

2.1.3.2.1. Le domaine de liaison d'un facteur de transcription à l'ADN

La séquence de liaison d'un facteur de transcription à l'ADN n'est pas fixe. Au contraire, elle peut permettre une variabilité au niveau de sa composition en nucléotide sans pour autant que la reconnaissance par son facteur ne soit affectée. Cette séquence peut varier pour avoir une taille plus en moins courte, avec une certaine variabilité dans la composition des nucléotides, et peut même comporter un/des espace(s) indéterminé(s). Ainsi, un facteur peut se lier à deux sites de composition très proche mais avec une affinité différente dépendamment des nucléotides qui les composent.

En résumé, un facteur de transcription ne possède pas une séquence unique mais un ensemble de motifs, sur lesquels, il peut se fixer pour lier l'ADN avec des affinités différentes. Cette caractéristique est l'un des facteurs qui participent à la complexité de la régulation génique. En effet, cette variabilité de l'affinité de liaison permet aux facteurs de transcription de moduler le degré de transcription d'une protéine. Par exemple, en se fixant à deux séquences semblables mais non identiques (en termes de leur composition en nucléotides), un même facteur de transcription ne donnera pas le même degré de production d'une protéine.

Notez qu'il existe d'autres facteurs qui peuvent aussi influencer sur l'accès des facteurs de transcriptions à l'ADN. Par exemple, les modifications d'histones (composants de base de la chromatine) par acétylation, phosphorylation et méthylation, jouent sur la rigidité de la chromatine afin d'autoriser l'accès à l'ADN (Berger, 2002).

2.1.3.2.2. Analyse bio-informatique des séquences cis-régulatrices

En bio-informatique, les séquences de fixation d'un facteur de transcription à l'ADN sont modélisées. Le modèle est un motif qui les décrit, appelé aussi la séquence consensus. Cette représentation des sites de fixation doit être la plus fidèle possible à l'ensemble de l'information contenue dans ces séquences.

La séquence consensus est construite à l'aide d'une matrice qui provient de l'alignement des séquences produites par les méthodes expérimentales (décrites ci-dessus). Ainsi, chaque motif, reconnu par le facteur de transcription, est rapporté et sert à calculer la matrice, dite de score de position spécifique (PSSM) (pour "a position specific scoring matrix"). Les lignes de cette dernière représentent les quatre bases possibles de l'ADN et les colonnes l'ordre séquentiel ou la position dans le motif. Ainsi, une entrée dans la PSSM donne le nombre observé d'occurrences de la base correspondant à cette position (voir tableau 1).

Tableau 1 : Matrice de calcul (matrice TRANSFAC F\$PHO4_01)

Residue\position	1	2	3	4	5	6	7	8	9	10	11	12
A	1	3	2	0	8	0	0	0	0	0	1	2
C	2	2	3	8	0	8	0	0	0	2	0	2
G	1	2	3	0	0	0	8	0	5	4	5	2
T	4	1	0	0	0	0	0	8	3	2	2	2
Sum	8	8	8	8	8	8	8	8	8	8	8	8

En se basant sur la matrice PSSM, une probabilité pour une séquence requête peut alors être calculée (voir tableau 2). Elle serait le produit des fréquences des bases des positions individuelles dans la séquence. Une valeur de probabilité dépassant un certain "cutoff" déterminera, par la suite, si c'est un match ou non (Kel, 2003). Il faut aussi noter que la séquence consensus est celle qui possède la plus haute valeur de probabilité. En effet, elle est composée des nucléotides avec les fréquences les plus

élevées à chaque position. Par conséquent, elle est la plus susceptible d'être reconnue par le facteur de transcription en question. Aussi, lorsqu'à une position donnée, les quatre bases partagent la même fréquence, la séquence est dite non spécifique pour cette position. La lettre *N* est alors assignée pour identifier cette caractéristique.

Tableau 2 : Matrice de fréquence corrigée (Hertz, 1999)

Pos	1	2	3	4	5	6	7	8	9	10	11	12
A	0.15	0.37	0.26	0.04	0.93	0.04	0.04	0.04	0.04	0.04	0.15	0.26
C	0.24	0.24	0.35	0.91	0.02	0.91	0.02	0.02	0.02	0.24	0.02	0.24
G	0.13	0.24	0.35	0.02	0.02	0.02	0.91	0.02	0.58	0.46	0.58	0.24
T	0.48	0.15	0.04	0.04	0.04	0.04	0.04	0.93	0.37	0.26	0.26	0.26
Sum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

$$f'_{i,j} = \frac{n_{i,j} + k/A}{\sum_{i=1}^A n_{i,j} + k}$$

Première option: le pseudo-poids distribué de façon identique

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^A n_{i,j} + k}$$

Deuxième option: le pseudo-poids en tenant compte de la priorité des résidus

- A* la taille de l'alphabet (= 4)
- n_{i,j}* les occurrences du résidu *i* à la position *j*
- p_i* la probabilité a priori du résidu *i*
- f_{i,j}* la fréquence corrigée des résidus *i* à la position *j*
- k* le pseudo-poids (arbitraire, 1 dans ce cas)

Notez qu'il existe aussi des matrices, appelées les matrices position-poids PWMs (pour Position Weight Matrix) (Schneider, 1986), qui indiquent le taux d'occurrence plutôt que le nombre d'occurrence. Plusieurs manières de convertir une matrice d'occurrence en matrice position-poids ont été développées. La plus simple est de diviser chaque colonne par la somme de chacune. Par contre, certaines, plus complexes, permettent d'utiliser les affinités du facteur de transcription pour les différents sites, tandis que d'autres sont basées sur les probabilités a priori d'observer telle base dans les séquences non codantes de l'organisme considéré. A titre d'exemple, nous présentons une de ces méthodes : la matrice de poids (modèle de Bernoulli) (Hertz, 1999) (voir table 3).

L'hypothèse de Bernoulli pour la matrice de poids: Si l'on suppose, pour le modèle du background et une succession indépendante de nucléotides (modèle de Bernoulli), que le poids W_s d'un segment de séquence S est simplement la somme des poids des nucléotides des positions successives de la matrice $(W_{i,j})$, alors, il est commode de convertir la PSSM en une matrice de pondération, qui peut ensuite être utilisée pour attribuer un score à chaque position d'une séquence donnée (une séquence requête).

Tableau 3 : Matrice de poids (modèle de Bernoulli) (Hertz, 1999)

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.79	0.13	-0.23	-2.20	1.05	-2.20	-2.20	-2.20	-2.20	-2.20	-0.79	-0.23
0.175	C	0.32	0.32	0.70	1.65	-2.20	1.65	-2.20	-2.20	-2.20	0.32	-2.20	0.32
0.175	G	-0.29	0.32	0.70	-2.20	-2.20	-2.20	1.65	-2.20	1.19	0.97	1.19	0.32
0.325	T	0.39	-0.79	-2.20	-2.20	-2.20	-2.20	-2.20	1.05	0.13	-0.23	-0.23	-0.23
1.000	Sum	-0.37	-0.02	-1.02	-4.94	-5.55	-4.94	-4.94	-5.55	-3.08	-1.13	-2.03	0.19

Les poids W_{ij} de la matrice position-poids, présentée sur le tableau 3, sont calculés en utilisant la formule suivante:

$$W_{ij} = \log \left| \frac{f'_{ij}}{p_i} \right|$$

$$f'_{ij} = \frac{n_{ij} + p_i k}{\sum_{i=1}^A n_{ij} + k}$$

où:

- A = taille de l'alphabet (4 pour l'ADN)
- n_{ij} = nombre d'occurrences du résidu i en position j de la matrice d'occurrences
- p_i = probabilité *a priori* du résidu i ($p_A=p_T \sim 0.325$ et $p_C=p_G \sim 0.175$)
- k = pseudo-poids (arbitraire, 1 dans notre exemple)
- f'_{ij} = fréquence relative du résidu i en position j , corrigée par le pseudo-poids k
- $W_{i,j}$ = poids du résidu i en position j

Le poids est positif, lorsque $f'_{i,j} > p_i$ (positions favorables à la liaison du facteur de transcription).

Le poids est négatif, lorsque $f'_{i,j} < p_i$ (positions non favorables à la liaison du facteur de transcription).

La matrice, ainsi produite, représente les sites de fixation de haute et moyenne affinité au facteur de transcription. Pour plus de détails sur cette méthode, le lecteur peut se référer à (Hertz, 1999).

Par ailleurs, il est important de noter que d'autres, comme dans (Foat, 2006; Roeder, 2007), ont construit leur modèle computationnel en se basant sur une approche biophysique. Celle-ci représente l'état lié et non lié de la combinaison d'un motif à son facteur de transcription, sous forme de formules tenant compte de plusieurs paramètres physico-chimiques, tels que la température, la concentration des composés, l'énergie libre de la liaison par mole (Foat, 2006) et l'affinité de liaison (Helge, 2007).

Contenu informatif des PWMs

Les valeurs représentant les fréquences de chaque nucléotide à chaque position dans une matrice PWMs sont peu informatives. Pour faire ressortir l'information contenue dans chacune de ces positions, un indicateur ou score *IC* (pour "Information Content"), indiquant le degré de conservation des bases à chaque site, est utilisé. Dans notre cas, le contenu informationnel d'une matrice est une mesure de la variabilité entre un site de fixation (modélisé par la matrice), et la séquence aléatoire attendue pour cet organisme (modèle du background). C'est un indicateur de la qualité informative de la matrice.

La dernière ligne du Tableau 4 (voir la ligne somme sur le Tableau 4) montre le contenu informationnel de chaque position. On peut remarquer que les positions centrales sont nettement plus informatives que celles aux extrémités (voir Tableau 4).

Tableau 4: Contenu informationnel (Hertz, 1999)

Prior	Pos	1	2	3	4	5	6	7	8	9	10	11	12
0.325	A	-0.12	0.05	-0.06	-0.08	0.97	-0.08	-0.08	-0.08	-0.08	-0.08	-0.12	-0.06
0.175	C	0.08	0.08	0.25	1.50	-0.04	1.50	-0.04	-0.04	-0.04	0.08	-0.04	0.08
0.175	G	-0.04	0.08	0.25	-0.04	-0.04	-0.04	1.50	-0.04	0.68	0.45	0.68	0.08
0.325	T	0.19	-0.12	-0.08	-0.08	-0.08	-0.08	-0.08	0.97	0.05	-0.06	-0.06	-0.06
1.000	Sum	0.11	0.09	0.36	1.29	0.80	1.29	1.29	0.80	0.61	0.39	0.47	0.04

Le calcul du contenu informationnel est fait en utilisant les formules suivantes :

$$f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^A n_{i,j} + k}$$

$$I_{i,j} = f'_{i,j} \ln \left(\frac{f'_{i,j}}{p_i} \right)$$

$$I_j = \sum_{i=1}^A I_{i,j}$$

$$I_{matrix} = \sum_{j=1}^w \sum_{i=1}^A I_{i,j}$$

où:

- w = la largeur de la matrice (=12),
- $I_{i,j}$ = l'information du résidu i à la position j ,
- A = taille de l'alphabet (4 pour l'ADN),
- n_{ij} = nombre d'occurrences du résidu i en position j de la matrice d'occurrences,
- p_i = probabilité *a priori* du résidu i ($pA=pT \sim 0.325$ et $pC=pG \sim 0.175$),
- k = pseudo-poids (arbitraire, 1 dans notre exemple),
- f'_{ij} = fréquence relative du résidu i en position j , corrigée par le pseudo-poids k ,
- $W_{i,j}$ = poids du résidu i en position j .

Le contenu informationnel varie de 0 à 2. Le contenu informationnel est de 0 lorsque la probabilité des quatre bases est de 0.25 (aucune base n'est favorisée ou élue). Par contre, le contenu informationnel est de 2 lorsque la probabilité pour un nucléotide est de 1 (dans ce cas, les autres ont une probabilité de 0). Par conséquent, les contenus informationnels les plus élevés s'interprètent comme correspondants aux positions des nucléotides hautement conservés, alors que les contenus informationnels avec les valeurs les plus basses correspondent aux positions les moins conservées (Schneider, 1986).

CHAPITRE III

ANALYSE DE LA RÉGULATION GÉNÉTIQUE : NOUVELLES APPROCHES

Le problème avec les anciennes approches d'analyse de la régulation génique est qu'elles ont l'inconvénient d'un nombre élevé de faux positifs. Les gènes prédits et les tests de liaison ne s'adaptent pas à un grand nombre de séquences candidates. Ces expériences restent laborieuses avec une efficacité discutée et demandent beaucoup de temps de traitement et d'analyse sans garantir un succès au final.

Face à ces limitations, l'orientation a été vers des approches expérimentales pour le mapping des interactions protéine-ADN à l'échelle du génome. Ainsi, les essais de l'immuno-précipitation de la chromatine (ChIP) (Kim, 2006), couplé avec les puces "genome tiling microarrays" (ChIP-chip) (Iyer, 2001; Ren, 2000) ou le séquençage (ChIP-Seq) (Barski, 2007; Johnson, 2007), ont été une révolution et sont devenus des techniques populaires pour l'identification des cistromes.

ChIP -Seq a émergé comme l'un des outils les plus prometteurs pour l'établissement du profil des sites de liaison protéine-ADN et des modifications de la chromatine à l'échelle du génome (Nix, 2008). Malheureusement, il n'y a eu que peu de recherches sur la comparaison entre ChIP-chip et ChIP-seq, mais en théorie, ChIP-Seq est une amélioration et remplace de plus en plus la méthode ChIP-chip. En effet, la possibilité d'observer les modifications directement dans les séquences des sites de liaison des facteurs de transcription a fait de ChIP-seq la méthode la plus utilisée pour les essais

d'interaction protéine-ADN (Furey, 2012; He, 2014) et l'analyse de la régulation de l'expression des gènes (Jiang, 2015a; Jiang, 2015b).

Bien que les premiers efforts de ChIP-Seq aient été limités par les coûts du séquençage à haut débit (Kim, 2006), d'énormes progrès ont été réalisés ces dernières années pour le développement du séquençage massivement parallèle de la prochaine génération. Des dizaines de millions de courts tags (25-50 bases) peuvent maintenant être séquencés simultanément à moins de 1% du coût des méthodes traditionnelles du séquençage Sanger. Les technologies telles que "Illumina's Solexa" ou "Applied Biosystems' SOLiDTM" ont fait de ChIP-Seq une alternative pratique et potentiellement supérieure à ChIP-chip (Barski, 2007; Johnson, 2007).

Dans ce chapitre, nous présentons l'expérience ChIP-seq. La description détaillée sera subdivisée en deux grandes parties. La première décrira l'expérience de l'immuno-précipitation de la chromatine (ChIP), les anciennes et les nouvelles méthodes de séquençage ainsi que les objectifs d'une expérience ChIP-Seq, ses avantages par rapport à une expérience ChIP-chip, ses faiblesses ainsi que ses considérations expérimentales. La deuxième partie, quant à elle, concernera la description du déroulement de l'analyse computationnelle des données issues d'une expérience ChIP-seq.

PARTIE I : ChIP-Seq

3.1. Immuno-précipitation de la chromatine (ChIP)

L'immuno-précipitation de la chromatine est la technique de choix pour l'étude de la distribution des protéines sur le génome (voir Figure 8). Cette technique requiert une première étape de fixation des cellules, généralement par le formaldéhyde. L'addition de ce composé (Jackson, 1978) crée des pontages covalents (cross-linking) dans les interactions protéine-protéine mais aussi dans celles protéine-ADN. La chromatine est par la suite extraite et fragmentée en courts brins (de 200 à 400 pb) et les fragments d'ADN, associés à la protéine étudiée, sont captés à l'aide d'un anticorps spécifique à cette dernière. Après précipitation des complexes ADN-protéine-anticorps, les brins d'ADN associés à la protéine seront récupérés en les séparant du surnageant (contenant l'ADN non associé à la protéine d'intérêt). À la fin de la procédure, la partie protéique des complexes [ADN-protéine-anticorps] subira une protéolyse (la protéine d'intérêt sera isolée en la séparant de l'ADN. Cette séparation se fait en reversant le cross-link ADN-protéine à l'aide du chauffage ou avec de l'ADN) pour ne conserver que les fragments d'ADN purifiés, où, la protéine s'est fixée.

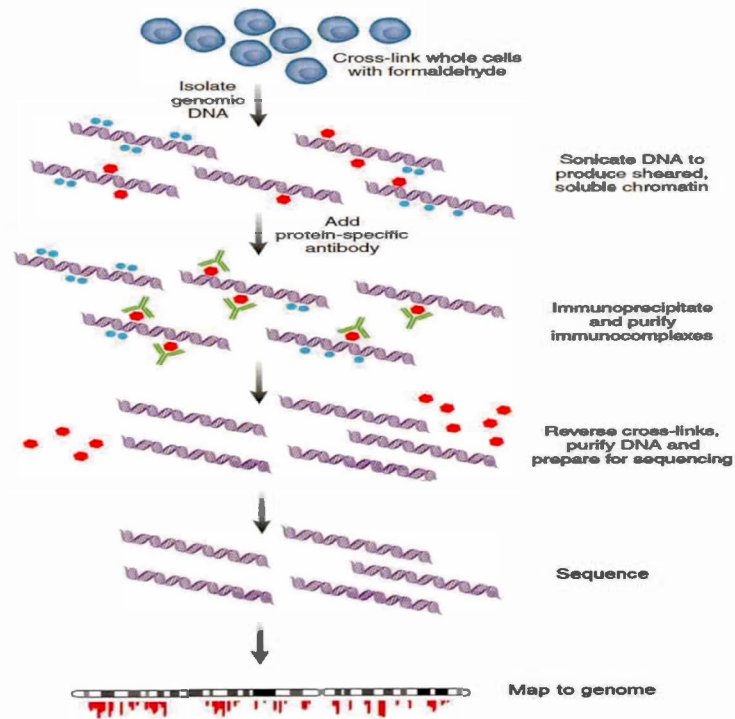


Figure 8 : Présentation d'une expérience ChIP-Seq. L'ADN et les protéines sont "cross-linked" et purifiés ; puis l'ADN lié est analysé par le séquençage massivement parallèle des courts reads (Mardis, 2007)

3.2. Identification des fragments issus de ChIP

Les fragments d'ADN obtenus à l'issue de ChIP peuvent être détectés par qPCR, hybridation sur des puces (*ChIP on chip*), ou plus récemment par le séquençage à haut débit (ChIP-Seq).

3.2.1. Hybridation sur des micropuces (ChIP on chip "Chromatin Immuno-Précipitation on chip")

Cette technique étudie les interactions protéine-ADN à l'échelle du génome en identifiant les fragments issus de ChIP par hybridation sur puces (microarray). Ces dernières sont constituées d'un ensemble de sondes d'oligonucléotides. Aujourd'hui, elles peuvent être conçues pour couvrir l'ensemble du génome, le cas des puces de très haute densité « tiling array », ou alors ciblent des régions bien précises, comme les promoteurs, des chromosomes spécifiques, ou des familles de gènes.

3.2.2. Immuno-précipitation de la chromatine suivie du séquençage haut débit (ChIP-seq)

Le développement de techniques très rapides des dernières années ont permis l'émergence de nouveaux outils de séquençage à haut-débit, constituant la famille des NGS ou Next-Generation Sequencing (voir la section NGS). Les NGS ont été appliqués dans de nombreux domaines du séquençage ou reséquençage des génomes entiers, le profil d'expression des gènes par séquençage des ARNm (RNA-seq), la caractérisation des sites hypersensibles à la DNase I, etc.

L'immunoprécipitation de la chromatine suivie du séquençage à haut-débit (ChIP-seq) a été une des premières applications des NGS. Les premières études ont été publiées en 2007 (Barski, 2007; Mikkelsen, 2007; Robertson, 2007). Dans cette technique, les fragments d'ADN issus de l'expérience ChIP sont séquencés directement au lieu d'être hybridés sur une puce.

3.3. Définition du séquençage

Le séquençage, ou la connaissance de la séquence d'ADN, consiste à définir la succession des quatre bases qui la compose à savoir l'adénine, cytosine, guanine, et thymine. Il a vu le jour à la fin des années 1970, soit une vingtaine d'années après la description de la structure de la molécule par Watson et Crick. Cette technique a valu le prix Nobel en 1980, a révolutionné la biologie moléculaire et elle est essentielle pour pratiquement toutes les branches de la recherche biologique.

3.4. Séquençage de la première génération

3.4.1. Le séquençage Maxam et Gilbert

L'histoire du séquençage a commencé avec Maxam et Gilbert, en 1977. La méthode est basée sur une modification chimique de l'ADN et le clivage subséquent à des bases spécifiques. Brièvement, l'ADN à séquencer est tout d'abord marqué en 5' avec phosphore 32 (dATP), puis clivé après l'acide nucléique A, G, C ou T par divers réactifs chimiques. Les fragments ainsi produits sont ensuite séparés par électrophorèse en gel de polyacrylamide. L'analyse de la radiographie correspondante permet de déterminer la séquence du brin analysé.

Malheureusement, cette méthode exige beaucoup d'ADN et la purification de fragments à plusieurs reprises. Les lectures sont relativement courtes sur des systèmes de gels manuels et l'automatisation n'est pas disponible (séquenceurs).

3.4.2. Le séquençage Sanger et son automatisation

Dans la même année, Frederick Sanger a développé une autre technologie de séquençage de l'ADN, basée sur la méthode de terminaison de chaîne. Elle repose sur une synthèse enzymatique des fragments d'ADN après leur amplification par clonage. Elle exploite la propriété que possèdent les ADN polymérases à synthétiser un brin complémentaire à partir d'un brin matrice en présence de dNTPs (désoxyribonucléotides triphosphate). Des ddNTPs (didésoxyribonucléotides triphosphate), marqués par fluorescence, sont ajoutés au milieu; ceux-ci servent de terminateurs aléatoires d'élongation dans la réaction. En faisant migrer sur gel de polyacrylamide, la séquence peut alors être reconstituée.

La méthodologie de séquençage Sanger n'exige que peu ou pas de purification des matrices d'ADN, aucune coupure avec des enzymes de restriction et aucun marquage de l'ADN à séquencer. Malheureusement, à cause de la faible processivité de l'ADN polymérase qui se détache au cours de la synthèse, elle reste limitée au séquençage d'un millier de nucléotides consécutifs.

En raison de son rendement élevé et de sa faible radioactivité, le séquençage Sanger a été adopté comme la principale technologie, de la «première génération», des applications de séquençage commerciales et des laboratoires (Sanger, 1975). Il a dominé l'industrie des génomes sur presque deux décennies et a permis aux scientifiques d'élucider l'information génétique à partir de n'importe quel système biologique donné et de faire de considérables réalisations, telles que le séquençage et l'assemblage du premier génome humain (Human Genome Sequencing Consortium, 2004). Cependant, malgré ces exploits, cette technologie a toujours été entravée par des limites inhérentes dans le débit, la vitesse et la résolution, qui ont eu des répercussions sur l'extraction de l'information essentielle.

3.5. Le séquençage haut débit (SHD)

3.5.1. Introduction

Pour surmonter les obstacles posés par le séquençage de la première génération, une technologie entièrement nouvelle, appelée Next-Generation Sequencing (NGS), a été nécessaire. Cette approche, totalement différente, a déclenché une révolution dans la science génomique en modifiant les procédés scientifiques et les applications biologiques.

Les NGS fournissent un débit plus élevé à une vitesse sans précédent par le séquençage de millions de courts fragments d'ADN en parallèle (Mardis, 2008; Metzker, 2010). Ils permettent de passer de plusieurs années de séquençage à quelques heures, pour produire 100 fois plus de *reads*, avec un coût 1000 fois plus réduit (von Bubnoff, 2008). Grâce à cette puissance de séquençage, il est maintenant possible de repousser les limites de la biologie actuelle, par exemple, étudier l'expression des gènes en caractérisant des transcrits restés inconnus, des épissages alternatifs et des variants de gènes (Myers, 2007; Wang, 2009). D'ailleurs, ces nouvelles technologies ont déjà permis des avancées importantes comme en génomique avec le projet de séquençage 1000 *génomomes* (Via, 2010) et en transcriptomique avec la détection de gènes de fusion dans les cancers (Maher, 2009).

3.5.2. Les plateformes NGS

Différentes plateformes existent à l'heure actuelle. Chacune a mis en place sa propre méthodologie de préparation des matrices, de séquençage, de capture d'image, et

d'analyse de données. Concentrons nous essentiellement sur les trois plates-formes les plus couramment utilisées dans la communauté scientifique : Roche 454 (introduit en 2005), Illumina (lancé en 2006) et ABI SOLiD (apparu en 2008). Les trois plates-formes séquencent l'ADN en mesurant et en analysant les signaux, émis au cours de la création du second brin d'ADN, mais diffèrent dans la manière dont le second brin est généré. Afin de produire des signaux détectables, l'ADN matrice est fragmenté en petits morceaux, amplifié et immobilisé sur une lame de verre avant le séquençage. Le tableau 5 présente un comparatif des performances des séquenceurs de nouvelle génération actuels et contient des informations, telles que le mécanisme de séquençage utilisé, les applications, les coûts, etc.

Tableau 5 : (a) Avantages et mécanismes des séquenceurs. (b) Composants et coût des séquenceurs. (c) Application des séquenceurs (Liu, 2012).

(a)

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

(b)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Instrument price	Instrument \$500,000, \$7000 per run	Instrument \$690,000, \$6000/(30x) human genome	Instrument \$495,000, \$15,000/100 Gb	Instrument \$95,000, about \$4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	\$10	\$0.07	\$0.13	\$2400

(c)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Resequencing		Yes	Yes	
<i>De novo</i>	Yes	Yes		Yes
Cancer	Yes	Yes	Yes	
Array	Yes	Yes	Yes	Yes
High GC sample	Yes	Yes	Yes	
Bacterial	Yes	Yes	Yes	
Large genome	Yes	Yes		
Mutation detection	Yes	Yes	Yes	Yes

3.5.2.1. Le système Roche 454

Roche 454 a été le premier système de la prochaine génération à succès commercial. Ce séquenceur implémente la technologie de pyroséquençage (<http://my454.com/products/technology.asp>), qui, au lieu d'utiliser des didésoxynucléotides pour terminer la chaîne d'amplification, repose sur la détection de la pyrophosphate, libérée lors de l'incorporation de nucléotides. La longueur de read est un caractère distingué de Roche par rapport aux autres systèmes NGS. Toutefois, son avantage le plus remarquable reste sa vitesse : Ça prend seulement 10 heures du début à la fin du séquençage (Liu, 2012). En ce qui concerne les inconvénients, comme cette technique déduit le nombre de nucléotides incorporés à partir de l'intensité du signal, le système rencontre des problèmes, liés aux taux d'erreur relativement élevés, lorsque des homopolymères étendus à plus de 8 pb sont séquencés (Margulies, 2005). Ceci complique l'identification de petites insertions et suppressions. Par ailleurs, le coût élevé des réactifs demeure le défi majeur de Roche 454.

3.5.2.2. Le système AB SOLiD

Le séquenceur SOLiD (pour "Sequencing by Oligo Ligation Detection") adopte la technologie de séquençage di-basique. Cette dernière repose sur le séquençage de ligature : elle analyse l'ADN par ligature des sondes di-basique, marquées par fluorescence au premier brin, nécessitant la lecture de chaque base deux fois. La longueur de read de SOLiD a été initialement de 35 pb et le résultat des données était de 3G par cycle. Grâce à la méthode de séquençage di-basique, SOLiD pourrait atteindre une précision élevée allant jusqu'à 99,85 % après filtrage.

À la fin de 2007, ABI a produit le premier système SOLiD et à la fin de 2010, le système de séquençage de SOLiD 5500xl est sorti. De SOLiD à SOLiD 5500xl, cinq mises à niveau ont été produites en seulement trois ans. Le SOLiD 5500xl a des longueurs de reads améliorées, une précision et des résultats de données correspondant à 85 pb, 99,99 %, et 30 G par run, respectivement. Un run complet pourrait être terminé en 7 jours.

Le coût de séquençage est d'environ 40×10^{-9} \$ par base, mais la longueur courte des reads et les reséquençages spécifiques aux applications demeurent ses principales lacunes (<http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html>).

En raison de la nature de l'approche SOLiD, les "Calls" identifiés ne sont pas stockés dans un nucléotide mais en espace couleur, une propriété qui doit être prise en compte dans les analyses en aval. En se référant à la matrice de codage di-basique, la séquence couleur d'origine peut être décodée pour obtenir la séquence de base si les types de base pour une quelconque position dans la séquence sont connus.

Ici, nous nous sommes focalisé sur le séquençage de type Illumina, comme c'est la plateforme qui a été utilisée dans notre étude. Pour obtenir plus d'informations quant aux autres technologies, le lecteur pourra se référer aux revues de (Metzker, 2010), (Zhang, 2011) et de (Liu, 2012).

3.5.2.3. Le système Illumina GA/HiSeq

Le séquençage Illumina est basée sur la méthode « Cyclic reversible terminator » (CRT). De façon générale, à l'aide de l'ADN polymérase, les quatre types de nucléotides (ddATP, ddGTP, ddCTP, ddTTP), contenant différents colorants fluorescents clivables et un groupe de blocage amovible, viennent, une base à la fois, compléter la matrice. Un nucléotide, auquel est fixé un fluorophore, subi une modification de façon à bloquer l'addition du nucléotide suivant. Après l'incorporation, les nucléotides non fixés sont éliminés par un lavage. L'image de la fluorescence de chacun des quatre fluorophores est ainsi capturée, ce qui permet de déterminer l'identité du nucléotide ajouté à une colonie donnée (le signal pourrait être capturé par un appareil de charge couplée (CCD)). À la fin, le terminateur et le fluorophore seront clivés et un nouveau lavage permettra de les éliminer avant de passer au cycle suivant. Les étapes détaillées du séquençage Illumina sont présentées dans la figure 9.

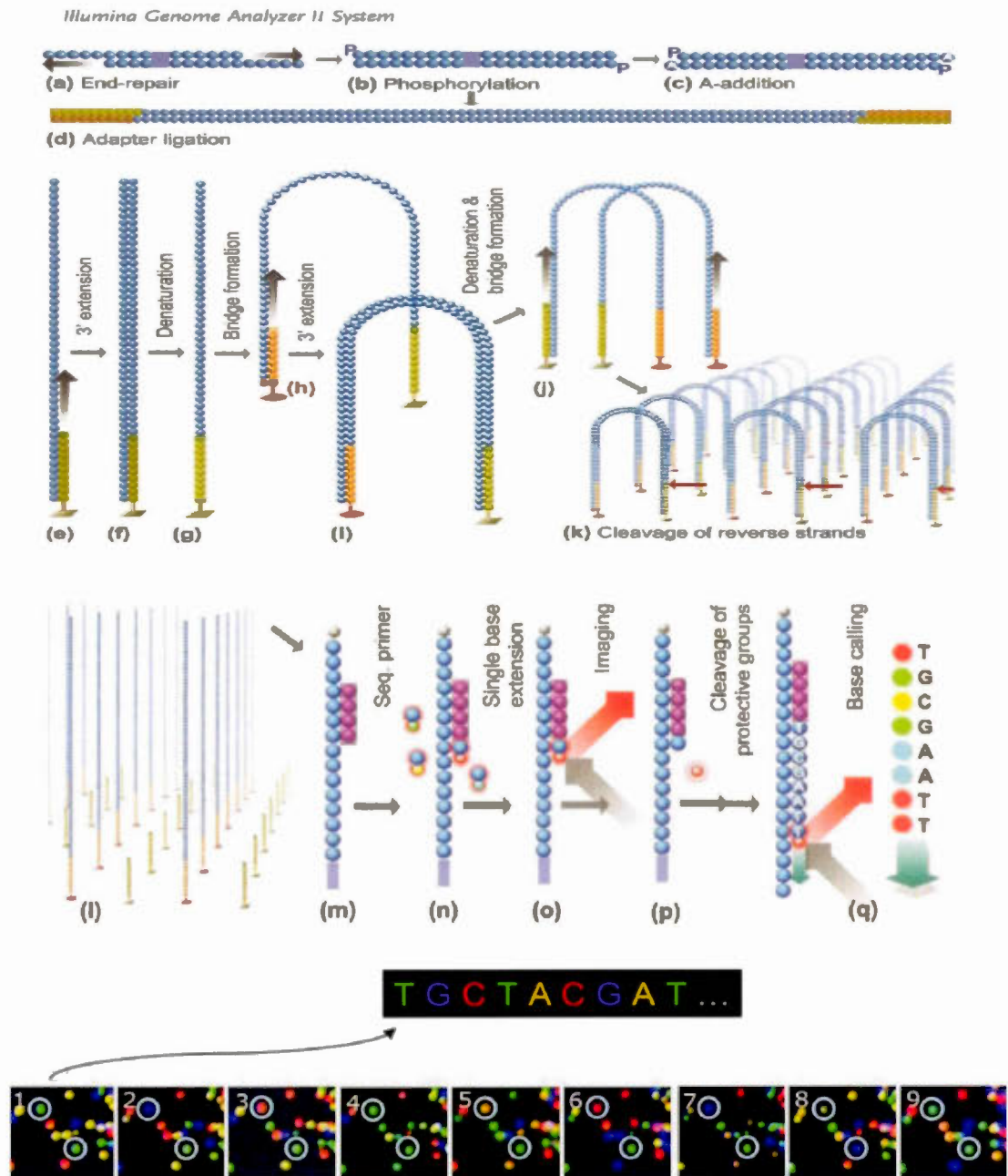


Figure 9 : Description du séquençage par Synthèse (Lakdawalla, 2008; Van Steenhouse, 2008).

Lors de la préparation de la librairie, les extrémités des fragments d'ADN sont réparées (a), phosphorylées (b), et une adénine est ajoutée (c). Enfin, des adaptateurs directs et inverses sont ligasés (d). La librairie avec les adaptateurs fixés est ensuite dénaturée en simples brins et greffée au « flow-cell » (e). La plupart des systèmes d'imagerie ne peuvent pas détecter un événement de fluorescence unique, par conséquent, les matrices d'ADN

doivent être amplifiées (pont d'amplification) pour former des clusters qui contiennent des fragments d'ADN clonal. L'amplification passe par deux phases principales: l'extension(f) et la dénaturation (g). L'ADN simple brin forme d'abord un pont en se liant aux amorces proches, liées à la « flow-cell » (h), ensuite s'étendu à nouveau et forme un double brin (i), subira à la fin une autre étape de dénaturation et de réhybridation pour former un nouveau pont (j). Ce processus est répété plusieurs fois jusqu'à la formation d'une colonie de fragments ADN identiques (k). Les brins inverses sont clivés, libérant l'extrémité 3' (k) qui sera par la suite bloquée. Enfin, les amorces du séquençage s'hybrident aux brins (m).

Avant le séquençage, la librairie est transformée en simples brins à l'aide de l'enzyme de linéarisation (Mardis, 2008), puis quatre types de nucléotides (ddATP, ddGTP, ddCTP, ddTTP) qui contiennent différents colorant fluorescent clivable et un groupe de blocage amovible viendraient compléter le template une base à la fois (n) grâce à l'ADN polymérase. Le signal pourrait être capturé par un (appareil de charge couplée) CCD (o). À la fin, le fluorophore et le terminateur seront clivés (p). Tout le processus est répété sur plusieurs cycles.

Actuellement, le nouveau Genome Analyser HiSeq 2000, lancé au début de 2010, peut séquencer des fragments de 100 pb et générer jusqu'à 200 giga bases par cycle d'utilisation (Lakdawalla, 2008). Par rapport à 454 et SOLiD, il est le moins cher avec 0,02 \$ / million de bases. De plus, avec l'aide des réactifs de Truseq v3 et des logiciels associés, HiSeq 2000 a beaucoup amélioré le séquençage à haute teneur en GC. MiSeq, un séquenceur lancé en 2011 et qui partage la plupart des technologies avec HiSeq, est particulièrement pratique pour amplicon et le séquençage des échantillons bactériens. Il pourrait séquencer 150 PE et générer 1.5 G /run en environ 10 heures, incluant le temps de la préparation de l'échantillon et de la librairie.

Plus d'informations sur les avantages et les inconvénients ainsi qu'un comparatif avec les systèmes HiSeq 2000 et AB SOLiD sont présentés dans les Tableaux 5 (a), (b) et (c).

3.5.3. Le choix du séquenceur

Concernant les trois systèmes NGS, décrits précédemment, Illumina HiSeq 2000 dispose du plus grand output et le plus bas coût de réactif, le système SOLiD a la plus

grande précision (Huse, 2007), et le système Roche 454 a la plus longue taille de lecture.

La quantité de données générées pour une expérience est de l'ordre de 450 Mpb pour Roche 454, 18-35 Gpb pour Illumina® et 30-50 Gpb pour SOLiD. La longueur respective des reads est, quant à elle, en moyenne de 330 pb pour le premier, 75-100 pb pour le second et 50 pb pour le dernier (Metzker, 2010). En résumé, Illumina® présente un bon compromis entre le nombre et la longueur des reads comparativement à SOLiD qui produit des reads trop courts et Roche 454® qui n'en produit pas assez.

En conclusion, le choix du séquenceur est souvent dépendant de l'application sous-jacente : nous avons la possibilité de choisir entre des séquences courtes ou longues, et entre des séquences peu ou fortement couvertes. Par exemple, Illumina® monopolise les applications liées à l'étude du transcriptome, car il propose un bon compromis entre la quantité et la longueur des reads, permettant ainsi une meilleure précision des jonctions des exons et l'expression des transcrits. Par contre, grâce à sa production de séquences longues, Roche 454® reste le leader pour les applications d'assemblage. Comme les très courtes séquences ne posent aucun problème dans les applications génomiques de type ChIP-Seq, du fait de l'absence d'épissage, d'une bonne couverture et que le nombre de reads qui couvrent la zone est très important, SOLiD est très apprécié.

3.6. Destin des fragments séquencés : objectifs d'une expérience ChIP-seq

Peu importe la plateforme de séquençage utilisée, une fois séquencés, les fragments issus de l'immunoprécipitation de la chromatine permettent de déterminer, *in vivo*, l'endroit où une protéine se lie au génome. La protéine en question peut être un

facteur de transcription, une enzyme de liaison à l'ADN, une histone, un chaperon ou un nucléosome. Autrement dit, l'objectif principal de l'expérience ChIP-seq est de prédire les régions du génome, où la protéine immuno-précipitée est liée, en trouvant les régions avec un nombre important de reads mappés (pics) (Bailey, 2013). En effet, la cartographie computationnelle des fragments séquencés identifie les emplacements génomiques de la protéine liée (FT dans notre étude), qui apparaissent sous forme de zones fortement séquencées, correspondant à une des extrémités des séquences d'ADN, collectées lors de l'immunoprécipitation de la chromatine. Les sites de liaison sont les zones situées entre deux de ces régions. L'identification de ces sites permet d'éclairer sur le rôle des interactions protéine-ADN dans l'expression génique et dans d'autres processus cellulaires.

3.7. Chip-seq vs ChIP on chip

3.7.1. Les avantages

La résolution de Chip-seq, à la base près, est sans aucun doute la plus importante amélioration par rapport à ChIP-chip. Les puces sont limitées en résolution en raison des contraintes et des incertitudes dans le processus d'hybridation. En effet, l'hybridation d'acides nucléiques est complexe et dépend de nombreux facteurs, y compris la teneur en GC et la longueur, la concentration et la structure secondaire des séquences cibles et des sondes. Ainsi, l'hybridation croisée entre des séquences imparfaitement appariées se produit fréquemment et contribue au bruit (Park, 2009). Un autre avantage de l'utilisation de ChIP-seq est la possibilité d'analyse des interactions inter-géniques à l'échelle du génome et le fait que la couverture du génome ne soit pas limitée par les séquences de sondes fixées sur la matrice. Ceci est

particulièrement important pour l'analyse de régions répétitives du génome, qui sont habituellement masquées sur les puces, comme dans le cas des génomes plus complexes, tels les mammifères, composés à plus de 50 % de séquences répétées. Le séquençage peut capturer les variations des séquences dans les répétitions qui pourront servir à aligner les lectures au génome. Les séquences uniques flanquant les répétitions sont également utiles pour l'alignement de ces séquences. Par exemple, seulement 48 % du génome humain est non répétitif, 80 % peut être cartographié avec des lectures de 30 pb et jusqu'à 89 % si les séquences de lectures sont de l'ordre de 70 pb (Rozowsky, 2009). En plus, bien que les sondes des puces à haute densité « tiling » puissent couvrir le génome entier, dans le cas des mammifères, un très grand nombre de puces est nécessaire pour accéder à l'ensemble de leur génome, ce qui multiplie les coûts (Kim, 2005). Enfin, ChIP-seq caractérise des cibles plus précises et donne une meilleure identification des motifs de liaison pour les interactions ADN - protéine.

En résumé, l'utilisation de ChIP-Seq offre la possibilité d'étudier le génome d'un individu à moindre coût, avec une meilleure couverture génomique, moins d'artefacts, et une plus haute résolution.

3.7.2. Les inconvénients

Comme toute autre technologie, ChIP-seq pose également des défis. Bien que les erreurs de séquençage aient été considérablement réduites, elles restent présentes, surtout vers les extrémités de chaque lecture. Ce problème peut être réglé par l'amélioration des algorithmes d'alignement (voir ci-dessous), qui n'aligneront que les premières bases de la séquence (seeds). Malheureusement, cette méthode n'est pas sans conséquences puisqu'elle peut causer une perte d'information. Même si des améliorations aient été apportées, il existe aussi un autre biais, lié à la forte teneur en

GC dans les fragments. Ceci a une influence à la fois lors de la préparation de la librairie mais aussi avant le séquençage, lors de l'amplification (Hillier, 2008; Quail, 2008). De plus, lorsqu'un nombre insuffisant de lectures est généré, il y a une perte de la sensibilité ou de la spécificité, sans oublier les problèmes techniques dans l'exécution de l'expérience, comme le chargement de la bonne quantité d'échantillon : trop peu d'échantillon se traduira par trop peu de tags ; trop d'échantillon se traduira par des marqueurs fluorescents trop proches les uns des autres, induisant des données de basse qualité.

3.8. Considérations expérimentales dans une expérience ChIP-seq

3.8.1. La qualité de l'anticorps

La qualité des données Chip-seq dépend essentiellement de la qualité de l'anticorps. Un anticorps spécifique et sensible donnera un niveau élevé d'enrichissement par rapport au background, ce qui facilite la détection des événements de liaison. De nombreux anticorps sont disponibles dans le commerce mais leur qualité est très variable. Malheureusement, une validation rigoureuse est un processus laborieux: pour les modifications d'histones, par exemple, la réactivité de l'anticorps avec les histones non modifiées ou des protéines non-histones doit être vérifiée par Western blot. En outre, la réactivité croisée avec des modifications d'histones similaires (dans le cas d'anticorps ciblant des modifications d'histones très proches, par exemple la di-vs la tri-méthylation dans le même résidu) doit être vérifiée à l'aide de deux anticorps indépendants, combinés avec l'ARNi contre les enzymes suspectées de déposer la modification, ou en utilisant la spectrométrie de masse du précipité peptidique (Park, 2009).

3.8.2. Quantité de l'échantillon

L'un des plus grand avantage de ChIP-Seq est la faible quantité d'échantillon nécessaire pour le séquençage. Une expérience ChIP typique nécessite de 10-100 ng d'ADN, correspondant à environ 10^7 cellules. Plusieurs protocoles de ChIP ont été développés afin de n'utiliser qu'un petit nombre de cellules, de 10^4 - 10^5 pour le profilage du génome (Acevedo, 2007), ou 10^2 - 10^3 pour la quantification par PCR au niveau de loci spécifiques (Dahl, 2008 ; Wu, 2009 ; O'Neill, 2006). Cependant, ces protocoles ne sont utilisés que pour certains facteurs de transcription ou des modifications d'histones abondantes (par exemple, l'ARN polymérase II ou l'histone H3K27me3), précipités par un anticorps de haute qualité. Pour la plateforme Illumina, 10 à 50 ng d'ADN sont recommandés, pouvant même descendre à 2 ng, tandis que le ChIP-chip requiert plus de 2 µg de matériel de départ. La quantité d'ADN et le nombre de cellules requis sont néanmoins dépendants de l'abondance de la chromatine associée au facteur ciblé et de la qualité de l'anticorps.

3.8.3. L'expérience contrôle

Les étapes expérimentales de l'expérience ChIP engendrent plusieurs sources d'artefacts potentielles. Par exemple, la sonication ou la digestion Mnase de l'ADN ne conduit pas à une fragmentation uniforme du génome. Les régions ouvertes de la chromatine ont tendance à être fragmentées plus facilement que les régions fermées, créant une distribution inégale des reads tout au long du génome. Les régions répétées semblent également être enrichies à cause du manque de précision du nombre de copies des répétitions dans les assemblages des génomes. Par conséquent, un pic dans le profil ChIP-Seq doit être comparé à la même région dans l'échantillon contrôle

correspondant afin de déterminer sa validité. En général, il y a trois choix couramment utilisés pour le contrôle : l'input ou l'ADN total avant l'immunoprécipitation, la IP « mock » (une immunoprécipitation réalisée sans anticorps), et une IP réalisée avec un anticorps non-spécifique (comme l'immunoglobuline G). Il n'existe pas un choix approprié d'un contrôle, chacun produit son lot d'artefacts, toutefois, l'input ADN a été utilisé dans presque toutes les études ChIP-Seq. Il corrige principalement des biais liés à la solubilité variable des différentes régions, la fragmentation de l'ADN et l'amplification.

Le problème avec IP mock est que très peu de matériel peut être immunoprécipité en l'absence d'un anticorps, ce qui conduit à de multiples IPs mock de mauvaise qualité. D'un autre côté, dans une série d'expériences, ChIP-chip par exemple, lorsque les données sont correctement normalisées, IP mock contribue peu au résultat global (Peng, 2007). De plus, pour les modifications d'histones, l'utilisation du ratio entre l'échantillon ChIP et les nucléosomes en vrac est également instructif. Dans ce cas, l'utilisation de ce ratio est largement suffisante puisqu'il correspond à la fraction de nucléosomes avec la modification particulière à cet endroit, pondéré sur toutes les cellules testées.

Une des difficultés pour une expérience contrôle ChIP-Seq est la quantité du séquençage nécessaire. Il est toujours important de séquencer en profondeur vue que les lectures se répartissent sur le génome en entier mais aussi parce que les biais seront plus difficiles à localiser si le séquençage n'est pas assez suffisant. À titre d'exemple, pour l'input ADN ou des nucléosomes en vrac, un grand nombre de tags séquencés serait réparti de manière uniforme sur le génome. Ainsi, pour obtenir des estimations précises le long du génome, un nombre important de tags sera nécessaire à chaque point, sinon, le fold enrichissement au niveau des pics va avoir de grandes

erreurs dues à un biais d'échantillonnage. De ce fait, le nombre total de tags à séquencer est potentiellement très important.

La distribution du bruit de fond est souvent déterminée empiriquement. Cependant, il peut être modélisé, par exemple suivant une loi de Poisson, à partir de l'échantillon lui-même (Mikkelsen et al, 2007). Enfin, quelque soit l'approche utilisée, il faut souligner que la distribution des lectures du bruit de fond n'est pas uniforme ni identique, selon le tissu ou le type cellulaire, et dépend du protocole et de l'expérience elle-même. Toutefois, il est possible d'éviter le séquençage d'un échantillon de contrôle si l'on ne s'intéresse qu'à des motifs de liaison différentiels entre les conditions ou entre des points dans le temps et lorsque la variation dans les préparations de la chromatine est petite.

3.8.4. Le séquençage paired-ends

Les fragments ChIP-seq sont généralement séquencés à une seule extrémité, l'extrémité 5', dans ce cas, le séquençage est dit « single-end ». Néanmoins, ils peuvent aussi être séquencés à leurs deux extrémités, ce qu'on appelle le séquençage « paired-end », comme c'est souvent le cas pour la détection des variations structurales dans le génome tels les insertions, les délétions et les larges réarrangements chromosomiques (Korbel, 2007). Le séquençage « paired-end » peut également être utilisé dans le ChIP-seq afin d'améliorer la spécificité lors de l'alignement des reads, en particulier dans les régions répétées, ou si l'on recherche des interactions à distance au niveau de la chromatine (Fullwood, 2010).

3.8.5. Le nombre de répliques

Des expériences replicats sont nécessaires pour assurer la reproductibilité des données. Pour les " microarrays ", les plates-formes et les protocoles se sont sensiblement améliorés afin que les replicats techniques des mêmes échantillons ne soient généralement plus à faire. Bien que cela soit susceptible d'être le cas pour ChIP-Seq (Marioni, 2008), les replicats biologiques sont toujours fortement recommandées pour tenir compte des variations entre les échantillons et pour vérifier la fiabilité des étapes expérimentales. En supposant qu'elles soient séquencées profondément, deux répliques concordantes seraient généralement suffisantes, une troisième réplique semble ajouter peu de valeur (Rozowsky, 2009).

PARTIE II : ANALYSE DES DONNÉES ISSUES D'UNE EXPÉRIENCE ChIP-SEQ

Dans cette partie, nous présentons le déroulement d'une analyse des données issues d'une expérience ChIP-seq. La description couvrira toutes les étapes dès la sortie des données de la machine de séquençage jusqu'à l'analyse finale, en passant par toutes les étapes : le contrôle de la qualité, l'alignement, le peak calling ainsi qu'une vue générale de la plupart des analyses en aval.

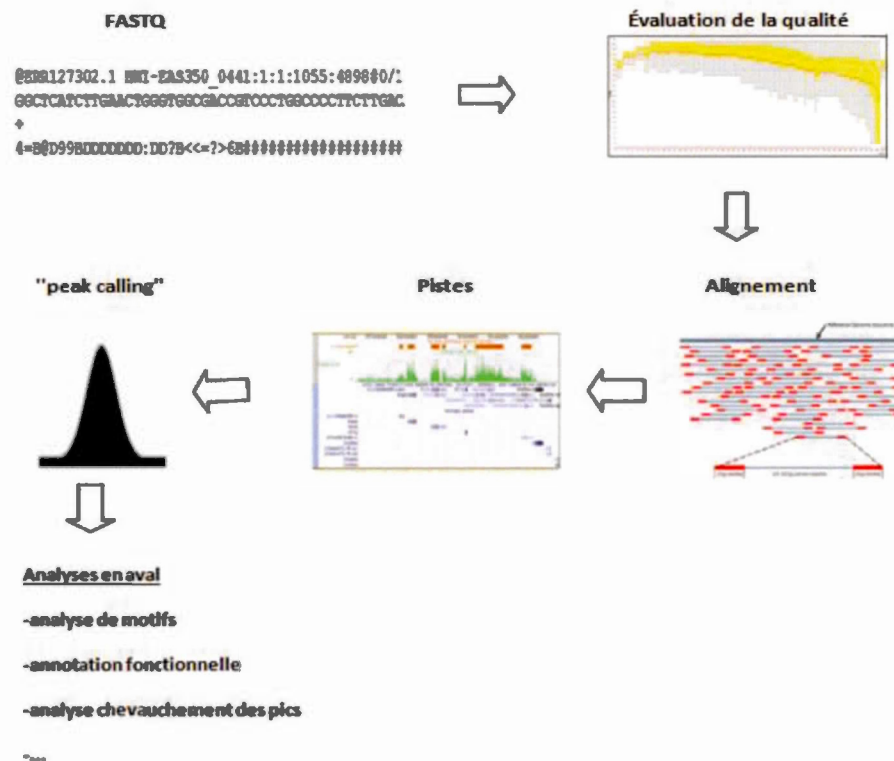


Figure 10 : Le workflow d'une expérience ChIP-Seq standard

Après l'achèvement des travaux de laboratoire et le séquençage réel, nous nous retrouvons confrontés à une énorme quantité de données brutes. L'analyse de ces données peut être décomposée en plusieurs étapes distinctes (voir Figure 10): l'évaluation de la qualité, l'alignement des reads à un génome de référence, le peak calling et l'analyse en aval. Cette dernière peut, à son tour, comporter plusieurs autres étapes en fonction des informations recherchées. Dans les paragraphes suivants, nous expliquerons, brièvement, chacune de ces étapes et nous passerons en revue les outils et logiciels disponibles, dont certains peuvent être retrouvés en suivant le lien (http://icbi.at/ngs_survey).

3.9. Contrôle de la qualité (Qc)

L'analyse des données issues du SHD nécessite une organisation judicieuse des *reads* afin d'assurer une détection précise des informations. En plus de la nécessité d'avoir une quantité suffisante de données, la qualité des reads doit aussi être vérifiée. En effet, les taux d'erreurs et les biais systématiques ont une grande influence sur leur traitement (Shendure, 2008). Ainsi, après avoir terminé le séquençage, la première étape de l'analyse est d'évaluer la qualité des reads bruts et de supprimer, de "trim" ou de corriger ceux qui ne respectent pas les normes et les standards définis.

Les données brutes, générées par les plates-formes de séquençage, sont compromises par des artefacts tels que les "base calling errors", les "indel", les reads de mauvaise qualité et la contamination par les adaptateurs (Dai, 2010). Ces erreurs sont très fréquentes et les plates-formes sont sensibles à un large éventail de défaillance chimique et instrumentale (Cox, 2010; Dohm, 2008). Récemment, les différents NGS ont été comparés sur leurs erreurs respectives (Suzuki, 2011). Cette analyse a montré

que les types d'erreurs sont différents selon la plateforme NGS et qu'il existe des biais qui leur sont intrinsèques. Par exemple, Illumina® produit un taux d'erreurs qui augmente avec la longueur du *read*, et il en produit davantage sur les nucléotides G et C.

Comme de nombreux outils d'analyse en aval ne sont pas en mesure de vérifier les reads de faible qualité, et afin d'éviter de tirer des conclusions biologiques erronées, il est nécessaire d'effectuer les tâches de filtrage et de trimming à l'avance. En général, ces mesures comprennent la visualisation des scores de qualité et les distributions de nucléotides, le trimming et le filtrage des reads. Ces tâches se basent sur le score de qualité des nucléotides et les propriétés des séquences, tels que les contaminations d'amorces, la teneur en N et le biais GC.

Plusieurs outils ont été développés pour effectuer ces différentes étapes de l'évaluation de la qualité (voir Tableau 6 de 11 outils sélectionnés). Par exemple, les outils autonomes NGSQC Toolkit (Dai, 2010) et PRINSEQ (Schmieder, 2011) sont en mesure de gérer les fichiers FASTQ et (SFF) 454, de produire des rapports sommaires, de filtrer et de trimmer des reads. De son côté, FastQC (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>) est compatible avec toutes les plateformes principales de séquençage et produit des graphiques et des tableaux récapitulatifs pour évaluer rapidement la qualité des données. Quant à l'outil ContEst (Cibulskis, 2011), il peut être utilisé pour estimer la quantité de contamination croisée inter-échantillon. En ce qui concerne Galaxy, ce dernier offre un outil intégré (Blankenberg, 2010) qui crée des statistiques sommaires FASTQ et effectue des tâches flexibles de trimming et de filtrage. De leur côté, htSeqTools (Planet, 2012) et SolexaQA (Cox, 2010) comprennent l'évaluation de la qualité, le traitement et la fonctionnalité de visualisation. En outre, d'autres logiciels ont été publiés mais ne supportent que la plate-forme Illumina comme par exemple, FASTX-Toolkit (http://han.nonlab.cshl.edu/fastx_toolkit), PIQA (Martinez-Alcantara, 2009) et TileQC (Dolan, 2008). D'autres fournissent certaines fonctionnalités spécialisées (exemple, TagCleaner (Schmieder, 2010).

Tableau 6 : Evaluation de la qualité et traitement de reads (Pabinger, 2013)

Nom	Système d'exploitation	Entrée	Sortie	Plateformes supportées	Rapport	Retrait de Tag (1)	Filtrage	"Trimming"
ContEST (cibulksis, 2001)	Lin, Mac, Win	BAM, VCF, FASTA	TXT	Illumina, ABI SOLiD, 454	non	non	non	non
FastQC (www.bioinformatics.babraham.ac.uk)	Lin, Mac, Win	(CS) FASTQ, SAM, BAM	HTML	Illumina, ABI SOLiD	oui	non	non	non
FASTX-Toolkit (hannonlab.cshl.edu)	Lin, Mac, web interface	FASTA, FASTQ	FASTA, FASTQ	Illumina	oui	oui	oui	oui
Galaxy (Blankenberg, 2010)	Lin, Mac, web interface, Cloud instance	FASTQ	FASTQ	Illumina	oui	oui	oui	oui
htSeqTools (Planet, 2012)	Lin, Mac, Win	FASTQ	Graphs	Illumina	oui	non	non	non
NGSQC (Dali, 2010)	Lin	FASTA (ref), FASTQ, CSFASTA, QUAL FASTA	HTML	Illumina, ABI SOLiD	oui	non	non	non
PIQA (Martinez-Alcantara, 2009)	Lin, Mac, Win	FASTQ, bustard, output, SCARF	HTML, TXT	Illumina	oui	non	non	non

PRINSEQ (Schmieder, 2011)	Lin, Mac, Win, web interface	FASTA, FASTQ, QUAL FASTA	FASTA, FASTQ, QUAL FASTA, HTML	Illumina, 454	oui	non	oui	oui
SolexaQA (Cock, 2010)	Lin, Mac	FASTQ	FASTQ, PNG	Illumina, 454	oui	non	non	oui
TagCleaner (Schmieder, 2010)	Lin, Mac, web interface	FASTA, FASTQ	FASTA	454	non	oui	non	non
TileQC (Dolan, 2008)	Lin, Mac	Eland output	Graphs	Illumina	oui	non	non	non

(1) Artifacts tels que adaptateurs, amorces ...

Après que les reads aient été traités et aient rencontrés une certaine qualité standard, ils sont habituellement alignés sur un génome de référence.

3.10. Alignement des courtes séquences de reads

Le problème d'alignement de la courte séquence de read est similaire au problème commun de correspondance de sous chaînes dans les systèmes de traitement des données. La comparaison d'une chaîne courte (read) à une longue chaîne de référence (génome de référence) peut donner un alignement dit valide ou non valide. La validité est mesurée en termes de mismatches ou gaps. Un mismatch se produit lorsqu'une base sur le génome de référence et une autre sur le read sont alignées mais sont différentes, tandis qu'un gap se produit lorsqu'une base est alignée avec un espace vide. L'un ou l'autre de ces alignements peut être la "bonne réponse" s'il permet à de

nombreuses autres paires de base, dans l'ensemble des reads, d'être alignées avec la séquence du génome de référence.

Généralement, il y a deux principales sources du génome de référence humain assemblé: l'université de Santa Cruz (UCSC) qui héberge aussi le répertoire central pour les données ENCODE (Rosenbloom, 2012), et "the Genome Reference Consortium (GRC)" qui se focalise sur la création d'assemblage de référence (reference assemblies) (<http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc.>). Les deux ressources fournissent plusieurs versions du génome humain: UCSC offre les versions hg18 et hg19 tandis que GRC fournit GRCh36 et GRCh37. Ensemble, ils constituent les génomes de référence les plus largement utilisés. Les assemblages des génomes humains UCSC (hg) et GRC (GRCh) sont identiques (<http://genome.ucsc.edu/FAQ/FAQreleases.html#release4.>) mais ils diffèrent du point de vue de leur nomenclature (exemple: UCSC utilise un préfixe "chr").

Ces dernières années, plusieurs programmes d'alignement ont été développés pour traiter efficacement des millions de courts reads (Yu, 2012). Le tableau 7, ci-dessous, est une liste des plus couramment utilisés.

La majorité de ces programmes peuvent être classés en deux catégories. Ceux qui utilisent l'indexation de table de hachage, comme dans le cas de BLAT (Kent, 2002) ou de SSAHA2 (Ning, 2001) et les algorithmes qui utilisent une sorte d'indexation d'arbre compressée basée sur la transformation Burrows-Wheeler (Burrows, 1994).

Tableau 7 : Table des logiciels d'alignement (Pabinger, 2013)

Nom	Système d'exploitation	Entrée	Sortie	plateformes Supportées	Méthode d'indexation	Alignement avec "gap"
BarraCUDA (Klus,2012)	Lin	FASTQ	SAM	Illumina	index FM (BWT)	oui
BFAST (Homer,2009)	Lin	FASTQ	SAM	Illumina, ABI SOLiD, 454	Multiple (hachage, arbre, ...)	oui
Bowtie (Langmead, 2009)	Lin, Mac, Win	FASTQ, FASTA	SAM	Illumina, ABI SOLiD	index FM (BWT)	non
Bowtie2 (Langmead, 2012)	Lin, Mac, Win	FASTQ, FASTA, QSEQ	SAM	Illumina, 454	index FM (BWT)	oui
BWA (Li,2009)	Lin	(CS)FASTQ, FASTA	SAM	Illumina, ABI SOLiD(1)	index FM (BWT)	oui
BWA-SW (Li,2010)	Lin	FASTQ, FASTA	SAM	454	index FM (BWT)	oui
ELAND (Cox,2007)	Lin	FASTQ, FASTA	SAM	Illumina	-	non
MAQ (Li,200)	Lin	FASTQ, FASTA	Maq	Illumina	basé sur le Hachage	oui
Mosaik (http://code.google.com/p/mosaik)	Lin, Mac, Win	FASTQ, FASTA	SAM, BED, plusieurs autres	Illumina, ABI SOLiD, 454	-	oui
mrFAST (Alkan,2009)	Lin	FASTQ, FASTA	SAM, DIVET	Illumina	basé sur le Hachage	oui

mrsFAST (Hach,201)	Lin	FASTQ, FASTA	SAM, DIVET	Illumina	basé sur le Hachage	non
Novoalign (http://novocraft.com)	Lin, Mac	FASTQ, (CS)FASTA	SAM, TXT	Illumina, ABI SOLiD	-	oui
SOAP2 (Li,2009)	Lin	FASTQ, FASTA	SOAP	Illumina	index FM (BWT)	oui
SOAP3 (Liu,2012)	Lin	FASTQ, FASTA	SAM	Illumina	index FM (BWT)	non
SSAHA2 (Ning,2001)	Lin, Mac	FASTA	SAM, GFF	Illumina, ABI SOLiD, 454	Indexation à arbres	oui
Stampy (Lunter,2011)	Lin, Mac	FASTQ, FASTA	SAM	Illumina, 454	index FM (BWT)	-
YOABS (Galinsky,2012)	Unix	-	-	Illumina	FM & Indexation à arbres	oui

3.10.1. Méthodes d'alignement basées sur le hachage

La première vague de programmes d'alignement, spécialement conçus pour l'alignement des courts reads issus des NGS, a été basée sur une structure de données de table de hachage pour indexer les données de séquences. De façon générale, une table de hachage est une structure de données commune capable d'indexer des données non séquentielles et complexes de manière à faciliter la recherche rapide (voir Figure 11). Cela est particulièrement approprié pour les séquences de reads, qui

sont extrêmement improbables de contenir toutes les combinaisons possibles de nucléotides et très susceptibles de contenir des doublons. Des exemples d'outils utilisant cette approche comprennent MAQ, SOAP et l'algorithme non publié ELAND propre à Illumina. Récemment, plusieurs d'autres sont apparus, tels que SHRiMP, ZOOM, BFAST ([http:// genome.ucla.edu/bfast /](http://genome.ucla.edu/bfast/)) et MOSAIK (<http://bioinformatics.bc.edu/marthlab/Mosaik/>) (voir tableau 7).

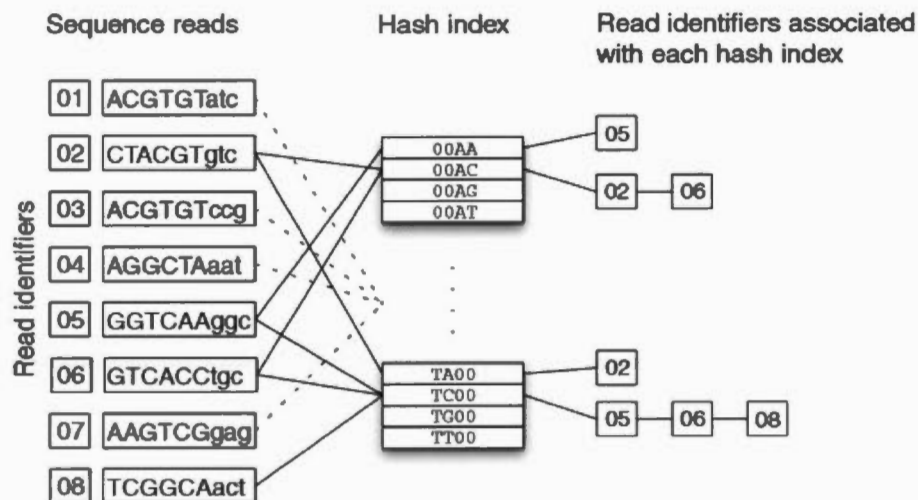


Figure 11: Schéma d'une stratégie d'alignement basée sur la table de hachage.

Des séquences reads avec les identifiants associés, les régions qui seront utilisées pour la sélection de seed en lettres majuscules et les seeds alignés de 0011 et 1100 sont présentés. Des identifiants de reads donnés sont associés aux seeds à l'aide d'une fonction de hachage (par exemple, une représentation d'entier unique de chaque seed). Une fois qu'une telle table de hachage a été construite pour le read input ou le génome de référence, les données correspondantes peuvent être scannées avec la même fonction de hachage, induisant comme résultat un sous-ensemble de reads beaucoup plus petit pour un alignement plus précis à chaque emplacement dans le génome (Flicek, 2009).

Les algorithmes basés sur les tables de hachage construisent leur table soit sur l'ensemble des reads input ou sur le génome de référence. Ils utilisent ensuite le génome de référence pour "scanner" analyser la table de hachage des reads input

(dans le premier cas) ou au contraire, ils utilisent l'ensemble de reads input pour "scanner" analyser la table de hachage du génome de référence (dans le second cas).

Chacune de ces deux méthodes a ses avantages et ses inconvénients. Par exemple, pour un ensemble donné de paramètres, les tables de hachage du génome de référence ont un besoin de mémoire constant, indépendamment de la taille de l'ensemble input de reads. Parfois, ce besoin peut être très important, en fonction de la taille et de la complexité du génome de référence. Par contre, les tables de hachage, basées sur l'ensemble input de reads, ont généralement des exigences mémoire petites et variables. En effet, elles dépendent du nombre et de la diversité de l'ensemble input de reads. Toutefois, lorsqu'il y a relativement peu de reads dans l'ensemble input, le temps de traitement pour analyser l'ensemble du génome de référence pourrait s'allonger. Parmi les algorithmes présentés dans le tableau 7, MAQ, ELAND, ZOOM et SHRiMP construisent une table de hachage des séquences de reads input, alors que SOAP, BFAST et MOSAIK hachent le génome de référence.

Quelle que soit la méthodologie de hachage, les algorithmes implémentent la table sous forme de «graines espacées» (se référer à la section 3.10.5 pour plus de détails sur la notion «graine»). Ces graines sont les régions de la séquence qui sont nécessaires pour avoir un modèle spécifique de matchs et de mismatches. Elles ont été popularisées pour l'alignement des séquences par le programme de "Pattern Hunter" (Ma, 2002). Une graine est de la forme 110 011, où 1 représente une position de la séquence qui est nécessaire pour le match, et le nombre de 1 désigne le «poids» de la graine. Par exemple, des 28 premières pb d'un read, le programme MAQ construit six tables de hachage correspondant aux graines de longueur 8 et de poids 4, puis analyse "scan" le génome de référence contre ces tables (Li, 2008). Cette technique assure que tous les hits avec deux mismatches puissent être trouvés, et plus de la moitié de ceux-ci avec trois mismatches. Vingt tables de hachage seraient nécessaires pour MAQ afin

de garantir que tous les reads avec trois mismatches soient trouvés. De son côté, ZOOM utilise des graines espacées, construites manuellement et de poids 14, pour permettre la détection jusqu'à quatre mismatches dans des reads de 50 pb (Ma, 2002). Par contre, SHRiMP utilise une approche q-gramme (Rasmussen, 2006) (se référer à la section 3.10.5 pour plus de détails sur la notion q-gramme). Pour qu'une région soit considérée comme un emplacement possible d'alignement, cette méthode exige que plusieurs graines espacées par read matchent (Rumble, 2009). Lors de son utilisation recommandée, MOSAIK hache toutes les positions dans le génome de référence et utilise une «base de données de saut» pour localiser efficacement l'information dans la table de hachage. Ainsi, il réduit les besoins en mémoire d'environ deux tiers par rapport à une implémentation naïve.

Une fois que les graines d'alignement ont été utilisées dans la création de la table de hachage et que les reads ont été associés à la région du génome (où ils sont le plus susceptibles de s'aligner), un algorithme spécialisé et précis, pourrait être utilisé pour déterminer l'emplacement exact de la séquence de read sur le génome de référence. De tels algorithmes comprennent les versions avec et sans " gap " de Smith-Waterman qui tirent parti des valeurs de qualité des bases des séquences.

3.10.2. Les méthodes de la transformation de Burrows-Wheeler

Récemment, une nouvelle génération de programmes d'alignement tels que BOWTIE (Langmead, 2009), BWA (Li, 2009) et SOAP2 (Li, 2009), basés sur la transformation Burrows-Wheeler (BWT) (Burrows, 1994), ont été développés.

La transformation Burrows-Wheeler est une technique des années 1990s développée à l'origine pour la compression de données (Langmead, 2009; Li, 2009). Les méthodes basées sur BWT utilisent une technique de matching de préfixes qui nécessite que quelques itérations pour produire un alignement valide. Ces méthodes utilisent généralement la structure de données d'index FM (un tableau de suffixes

comprimés), proposée par Ferragina et Manzini. Ces derniers ont introduit le concept, selon lequel, un tableau des suffixes est beaucoup plus efficace, s'il est créé à partir de la séquence de BWT, plutôt qu'à partir de la séquence d'origine (Ferragina, 2000). L'indice de FM conserve la capacité de la matrice de suffixe pour la recherche rapide de la sous-séquence. Il est souvent de la même taille ou de taille plus petite que la taille du génome d'entrée (Gräf, 2007). Par exemple, l'indice final pour le génome humain, utilisé à la fois par BWA et BOWTIE, est d'environ 2,3 Go en taille (Langmead, 2009 ; Li, 2009), tandis que SOAP2 utilise une routine différente qui mène à un indice final qui nécessite 5,4 Go (Li, 2009).

La création de la structure de données sous-jacente nécessite deux étapes. Dans la première étape, l'ordre de la séquence du génome de référence est modifié en utilisant la BWT. Ce processus réversible (à savoir, la séquence du génome initiale peut facilement être reconstituée) réorganise le génome de telle sorte que les séquences qui existent plusieurs fois apparaissent ensemble dans la structure de données (voir Figure 12). Ensuite, l'indice final est créé et utilisé pour l'alignement rapide des lectures sur le génome. La création de l'indice final peut être une étape gourmande en mémoire, toutefois, des méthodes existent pour créer l'indice dans relativement peu de mémoire au prix de plus de temps de traitement (Kärkkäinen, 2007).

Comme dans les méthodes basées sur le hachage, une fois que les reads ont été associés à la région du génome, où ils sont les plus susceptibles de s'aligner, des algorithmes plus sensibles peuvent être utilisés pour le résultat final de l'alignement.

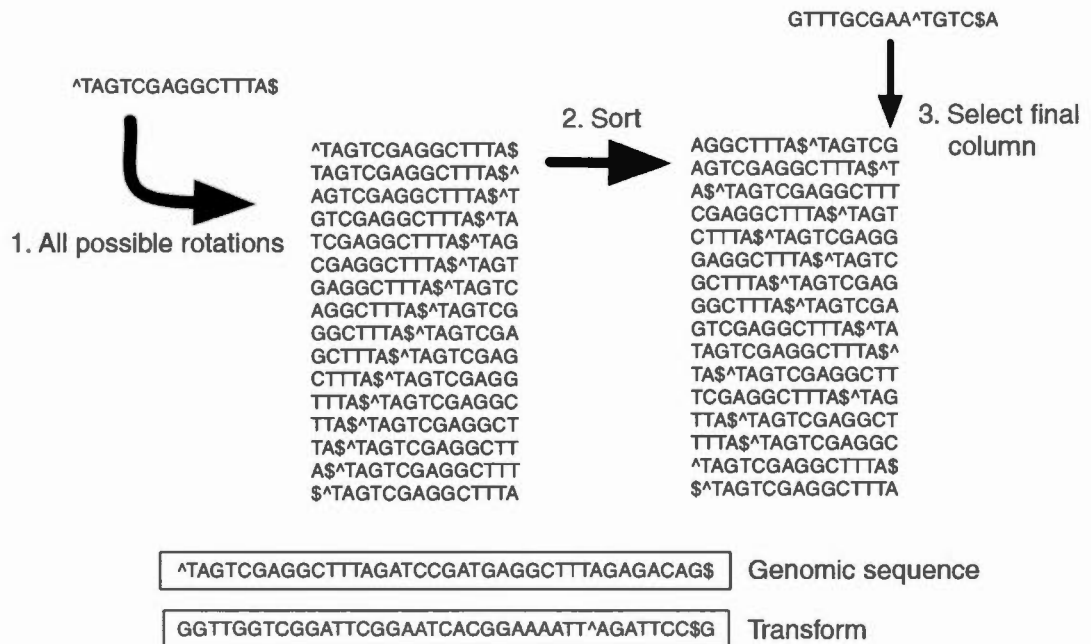


Figure 12 : La transformation de Burrows-Wheeler pour les données des séquences génomiques.

Pour créer une BWT d'une séquence génomique de 14-mer, on doit d'abord noter les points de début et de fin de la séquence puis construire toutes les rotations de la séquence donnée en prenant le premier caractère de la séquence et de le placer à la fin de la séquence (étape 1). Les caractères ^ et \$ marquent le début et la fin de la séquence, respectivement. Une fois que ces séquences sont créées, elles sont triées (étape 2). A partir de cette matrice de classement, la dernière colonne est sélectionnée comme séquence transformée (étape 3). Les séquences transformées ont exactement la même longueur et ont exactement les mêmes caractères que la séquence d'origine, mais dans un ordre différent. La séquence en bas est une séquence plus longue avec le même 14-mer qui démontre l'effet de l'utilisation de la séquence transformée sur une séquence input plus longue (Flicek, 2009).

3.10.3. Les tables de hachages vs BWT (*MAQ* vs *Bowtie*)

Au même niveau de sensibilité que leurs homologues à base de hachage, les implémentations de BWT sont beaucoup plus rapides. BWT est susceptible d'être la plus appropriée pour l'alignement des reads de ChIP-seq ou des applications similaires. Un autre avantage pour cette méthode est la capacité de stocker l'index

complet du génome de référence sur le disque et de le charger complètement en mémoire sur presque tous les clusters de calculs bio-informatiques standards (Flicek ,2009).

Bowtie est la méthode leader de l'état de l'art qui a mis l'accent sur l'utilisation des propriétés de la transformation de Burrows-Wheeler (BWT) pour indexer la séquence de référence (Figure 13). Un indexe basé sur BWT a une petite empreinte mémoire. Cet avantage a rendu Bowtie possible sur des ordinateurs avec seulement 2 GO de mémoire. Toutefois, la mémoire devient de plus en plus moins chère (alors que la taille du génome est constante), ce qui rend inutile de placer de telles restrictions de mémoire sur un logiciel d'alignement. L'augmentation de vitesse de 30 fois pour BOWTIE par rapport à MAQ (basé sur le hachage) est un exemple des augmentations de vitesse pour les reads single-end. Cette augmentation de vitesse est au prix d'une légère perte de la sensibilité de l'alignement (Rasmussen, 2006) (Figure 13).

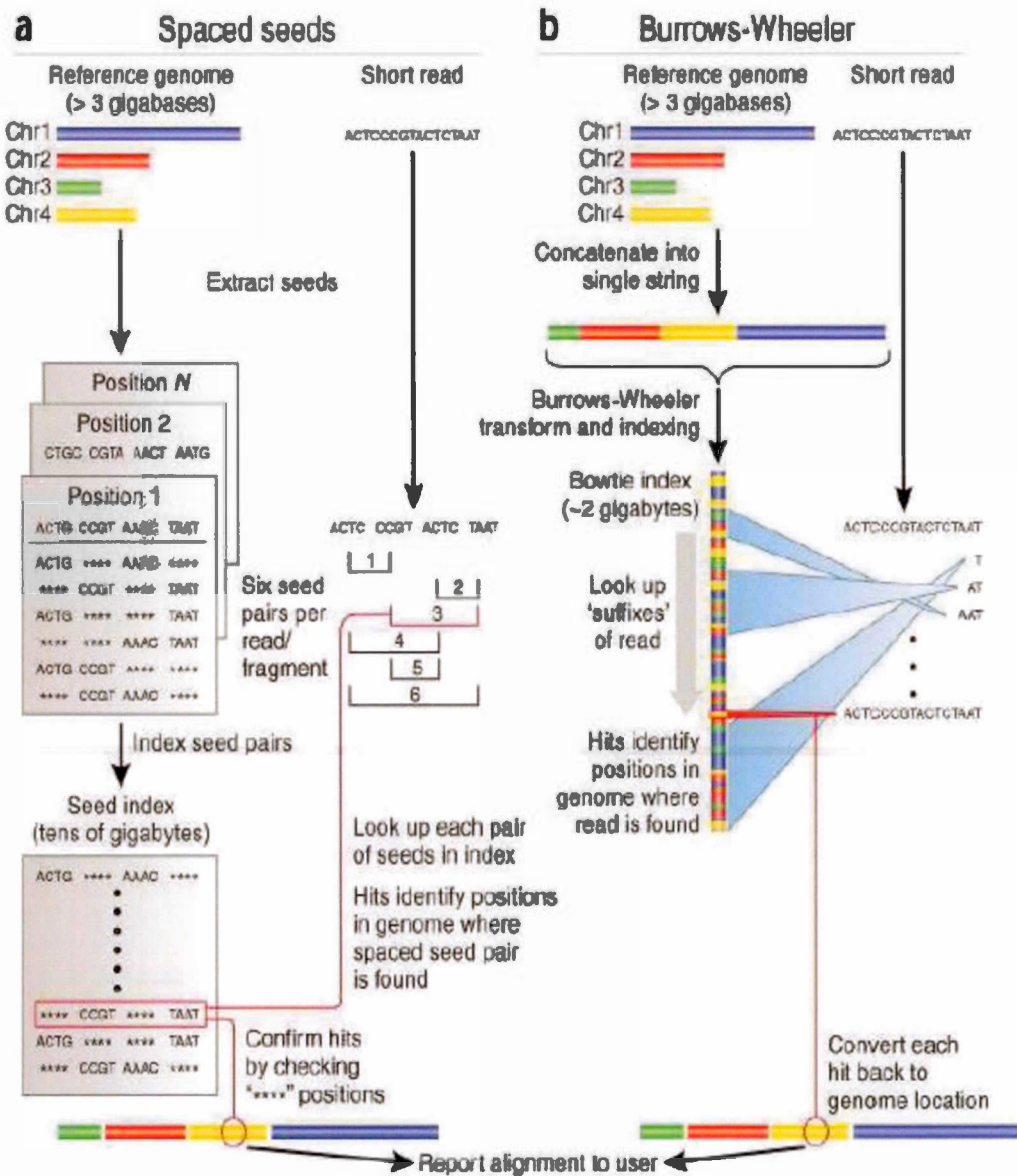


Figure 13: Algorithmes de *MAQ* (Spaced seeds) et *Bowtie* (Burrows-Wheeler) (Trapnell, 2009)

Avec la capacité d'exploiter les reads paired-end et pour une sensibilité similaire aux méthodes antérieures à base de hachage, l'augmentation de vitesse pour quelconque des programmes actuels basés sur BWT sera généralement de dix fois "tenfold" (Langmead, 2009 ; Li, 2009).

Les limites des méthodes basées sur BWT sont un autre exemple de compromis entre la vitesse et la sensibilité. En effet, la méthode BWT fonctionne bien avec les courts reads, mais ses performances se dégradent rapidement au fur et à mesure que la longueur des reads augmente. Par exemple, BWA est capable de trouver des alignements seulement dans une certaine distance d'édition de la séquence dans le génome de référence (Li, 2009). La distance d'édition est, de façon formelle, le nombre d'opérations nécessaires pour transformer une séquence à une autre, ce qui, dans le cas de l'alignement de séquence, est le plus souvent des gaps ou des mismatches. Cela limite efficacement le nombre combiné de mismatches ou de gaps qui peuvent être alignés (pour des reads de 100 pb, BWA permet 5 éditions; moins pour les reads plus courts et plus pour les reads plus longs). Comme le séquençage devient de plus en plus précis, cette limitation est susceptible de devenir moins importante pour les espèces avec des taux de polymorphisme relativement faibles, comme l'humain où presque tous les reads s'aligneront dans la distance d'édition.

3.10.4. Outils d'alignement pour les séquences courtes

3.10.4.1. Méthodes basées sur l'indexation des reads

Ces méthodes traitent les reads avant de les aligner sur le génome de référence.

- ELAND

D'abord, il y a des outils qui associent les reads à des graines espacées (voir dans la section 3.10.5), tel que le logiciel ELAND de ILLUMINA ou SeqMaq (Jiang, 2008). Une graine est une portion consécutive d'un read qui se trouve sur le génome (plus de détails dans la section 3.10.5), sachant que des différences peuvent exister entre les deux. Ces graines sont par la suite indexées dans une table de hachage. Ce genre d'algorithmes est plus économe et plus rapide que les alignements classiques (dans BLAT et BLAST par exemple) du fait que le génome est haché à la volée.

- Graines simples: Blat et Blast

L'indexation des séquences dans ces logiciels est basée sur des graines simples (voir dans la section 3.10.5) en utilisant le paradigme Seed and Extend et des alignements dynamiques du type Smith-Waterman (Kent, 2002 ; Altschul, 1990). Afin de pouvoir estimer la probabilité qu'une séquence approximative identifie la vraie localisation du read sur le génome, ces aligneurs se basent sur une étude statistique très complexe. Toutefois, ils ne tolèrent pas plus de deux substitutions et les séquences ne doivent pas dépasser 32 pb.

- MAQ

Le logiciel MAQ garantit l'alignement des 28 premières bases des reads avec au maximum deux substitutions, le reste de la séquence étant extrapolé. Il présente plusieurs améliorations par rapport aux méthodes précédentes en supportant des séquences plus longues (jusqu'à 63pb) et en intégrant une phase supplémentaire de vérification après l'alignement. Cette phase a pour but de filtrer les séquences potentielles par un score. Pour ce faire, il calcule la moyenne de la qualité des

substitutions dans la partie extrapolée, ainsi, il limite les artéfacts dus aux erreurs de séquences.

3.10.4.2. Méthodes basées sur l'indexation du génome

Contrairement à la méthode précédente, les reads sont recherchés sur un génome indexé. Ceci permet de résoudre le problème de saturation de la mémoire posé par la méthode précédente. En effet, dans cette dernière, la quantité de mémoire utilisée dépend directement de la quantité de reads à indexer. Comme les SHD produisent des quantités gigantesques de reads et que les tables de hachage sont gourmandes en mémoire, la solution inverse de l'indexation du génome s'avère nécessaire. L'idée originale provient d'une équipe (Ning, 2001), qui a adapté un index pour une longue séquence d'ADN en la découpant en petites portions continues mais suffisamment grandes de taille k pour créer le logiciel SSAHA (Ning, 2001). Cette idée d'indexer le génome plutôt que les reads, en utilisant des tables de hachage et en consultant les séquences à la volée, a été reprise par (Li, 2008) dans SOAPv1.

Cette approche inversée est cinq fois plus rapide que MAQ et donne des résultats similaires en termes de prédiction pour des séquences ≤ 60 pb. Toutefois, la table de hachage du génome occupe un espace mémoire important pouvant aller jusqu'à 20 fois l'espace utilisé pour le texte original. Afin de pouvoir appliquer l'idée, la solution était d'indexer chacun des chromosomes de façon indépendante.

Une méthode d'alignement de courts reads, tenant compte de substitution et des indels, a été présenté en 2009 (Hoffmann, 2009). Ce travail est basé sur l'utilisation d'arbres des suffixes pour économiser le temps de calcul. En effet, contrairement aux

tables de hachage, où chaque copie de la séquence est traité différemment, les arbres de suffixes identifient de façon plus précise les portions communes entre reads et génome et les regroupe dans un même endroit. Malheureusement, le coût en mémoire est important du fait du stockage d'avantage d'informations au niveau de chaque nœud, telles que les adresses du père et des fils. Dans la méthode d'Hoffmann et al, 2009, les chromosomes ont été traités de façon indépendante afin d'économiser l'espace mémoire. Ceci a des limitations sur certaines applications telles que la recherche de gène de fusion. Pour une utilisation moins coûteuse en mémoire, des travaux (Langmead, 2009 ; Li, 2009) se sont basés sur l'indexation compressée (Salson, 2009). Ainsi, le génome sera traité en un seul bloc et la lecture se fera sans décompression. Ces structures de compression ont fait de BOWTIE une méthode rapide tout en demeurant économe sur la mémoire. Elles ont par conséquent été adopté dans la version SOAPv2 (Li, 2009b ; Li, 2009] ainsi que dans BWA (Li, 2009) pour des séquences courtes (jusqu'à environ 100 pb), avec quelques substitutions et un gap, tandis que BWA-SW (Li, 2010) est efficace sur des séquences plus longues (≥ 150 pb) et est basé sur Smith-Waterman-like.

3.10.5. Outils d'alignement pour les séquences longues

3.10.5.1. Alignement approché pour les reads plus longs

Souvent, les aligneurs des courts reads de la première génération utilisent deux mismatches et aucun gap comme définition d'un alignement valide. Aujourd'hui, ces modèles sont devenus plus sophistiqués avec des programmes améliorés qui peuvent traiter des reads plus longs et avec des gaps. En effet, les avancées technologiques avec les machines de séquençage next-génération ont augmenté les longueurs de reads afin de produire ce qu'on appelle " les longs reads ". Cette tendance implique que les aligneurs basés sur BWT existants deviendront probablement, du point de vue

informatique, infaisables dans un avenir proche. En outre, les aligneurs de séquence auront également besoin de recevoir plus d'erreurs dans les modèles de correspondance (matching) (pour les reads longs), et la performance des méthodes existantes se détériorent rapidement lors de l'utilisation de ces modèles plus riches. Il faut donc dépasser les limites des outils qui n'autorisent qu'un petit nombre et types de différences (par exemple, une ou deux substitutions pour Bowtie) en développant des méthodes performantes capables de détecter des indels plus longs.

Comme mentionné auparavant, l'alignement de *reads* sur un génome de référence peut être vu comme une recherche de motif approchée (read) dans un texte (génome). Les principes de filtrations rapides, déjà adoptés par toutes les méthodes de recherche de motifs approchés, peuvent donc être exploités afin de trouver des stratégies de positionnement efficaces. Ces principes comportent deux étapes importantes : l'étape de filtration et celle de la vérification. La filtration a pour objectif d'éliminer des positions du texte ne pouvant présenter de similarités avec le motif. Ceci se fait par l'évaluation d'un critère simple. L'étape de la vérification, quant à elle, examine parmi les positions sélectionnées celles où débute une occurrence approchée du motif. L'efficacité de cette méthode de recherche dépend essentiellement de la capacité de l'étape de filtration. En effet, cette dernière peut, selon le critère de filtration, rater de vraies occurrences. On parle alors de filtration avec perte, et vis versa. En général, le critère de filtration est de complexité linéaire, voire sous-linéaire, tandis que la vérification peut être de complexité quadratique.

Il existe plusieurs méthodes de filtration (Pevzner, 1995) très largement utilisées en bio-informatiques, notamment dans BLAST, FASTA, BLAT, etc (Kent, 2002). Parmi elles, on trouve le lemme des q -grams (Owolabi, 1988; Jokinen, 1991), qui a été utilisé avec succès pour la recherche multiple de similarités dans l'alignement de

reads (par exemple dans QUASAR (Burkhardt, 1999)), et les graines espacées (Ma, 2002; Li, 2003; Burkhardt, 2003).

– Le filtre des q -grams

Le filtre des q -grams permet de chercher des occurrences approchées d'un motif dans un texte. Le lemme du q -gram dit que dans une occurrence approchée renfermant $\leq p$ erreurs, il existe au moins un sous-mot exact de longueur $\lfloor m - p + 1/q + 1 \rfloor$ (où m est la longueur du motif, p est le nombre d'erreur et q est la longueur de la fenêtre). Le pire cas est celui où les erreurs sont distribuées de façon régulière. Le sous-mot compris entre deux erreurs est identique entre le motif et le texte. La contrainte imposée par le lemme est que tous les matchs (correspondances) soient adjacents dans cette fenêtre de longueur q . Par exemple, pour un *read* de 50 *pb* qui contient 3 erreurs de séquence, il n'y a pas de portions plus petites que 12 *pb* ($m=50$, $p=3$ et $q=3$, alors $\lfloor 50 - 3 + 1/3 + 1 \rfloor = 12$) qui ne sont pas localisées, par conséquent les graines ont une taille de 12 au maximum. Ce critère de filtration, appelé *seed and extend*, peut être utilisé en combinaison avec l'indexation de tous les q -grams ou k -mers du génome, comme dans le cas de GASSST pour le traitement de *reads* génomiques et GSNAP pour les *reads* transcriptomiques (Rizk, 2010; Wu, 2010).

D'autres approches utilisent ce principe en l'appliquant sur les reads plutôt que sur le génome tel que dans RazerS (Rumble, 2009; Weese, 2009). En spécifiant un seuil comme paramètre du programme, RazerS ne garde que les alignements possédant un nombre suffisant de q -matchs. Par la suite, les meilleurs alignements sont conservés grâce à une analyse effectuée en se servant de vecteurs de bits (Myers, 1999).

– Les graines espacées

Les outils d'alignement tels que BLAST, ont déjà utilisé la méthode de graines contiguës pour la recherche d'un match potentiel. Le constat est que des mismatches

entre les séquences peuvent exister de sorte qu'on peut avoir des alignements mais aucune graine contigue suffisamment longue ne peut être alignée. Il fallait donc penser à assouplir la contrainte de contiguïté. C'est ainsi que sont apparues les graines espacées dont l'idée est d'améliorer les outils de mapping (Burkhardt, 2003).

Une graine espacée peut être définie par une fenêtre de longueur $w > q$ dans laquelle on obstrue quelques positions (*i.e.*, $w - q$). Cette idée de disperser des similarités sur une plus grande largeur de fenêtre a permis d'augmenter la sensibilité de la filtration. Les graines espacées sont utilisées avec succès comme dans ZOOM et PerM (Lin, 2008; Chen, 2009). Grâce à elles, la dépendance entre positions erronées dans les *reads* et les dépendances entre positions dans le codage SOLiD des *reads* a pu être modélisée, tels que dans le logiciel storm (Noé, 2010). Malheureusement, le prétraitement pour la conception des graines, incluant le choix de la longueur des graines ainsi que les positions à obstruer, est de complexité exponentielle, problème NP difficile (Kucherov, 2006; Ma, 2007; Nicolas, 2008). En plus, l'indexation des graines espacées est plus difficile et gourmande que celle de simples k -mers. Dans le cas de ZOOM, la conception des graines dépend des paramètres de l'alignement, ainsi l'utilisateur doit changer de graines à chaque fois que les paramètres changent (Rivals, 2009), ce qui rend son utilisation peu souple.

3.10.6. Le choix d'un logiciel d'alignement

Le choix d'un logiciel d'alignement approprié dépend de la plateforme de séquençage, le besoin de vitesse, et les ressources machine. Par exemple, si les données proviennent de la plateforme SOLiD, les outils supportant l'espace couleur "colorspace" sont des choix idéaux, tel que BWA. Si la vitesse est la considération majeur, Bowtie est à privilégier (voir Tableau 7).

Pour plus de lecture, voir de Bao et al pour plus de détails sur la comparaison des aligneurs (Bao, 2011).

En plus de la sélection du programme d'alignement, il faut tenir compte de trois remarques importantes:

- Pour surmonter le problème de l'ambiguïté lors de l'alignement des courts reads à un génome de référence, les reads paired-end se sont révélés être une solution utile et sont fortement recommandés, si ce n'est une exigence pour le séquençage d'exome entier et le séquençage de l'ensemble du génome.
- Comme les technologies NGS actuelles intègrent les étapes PCR dans leurs préparations de librairies, plusieurs reads originaires d'un seul modèle "template" pourraient être séquencés, interférant ainsi avec les statistiques de "variant calling". Pour cette raison, il est de pratique courante de supprimer les duplications PCR après l'alignement sur l'ensemble du génome.
- Afin de simplifier l'analyse, habituellement, seuls les reads alignés à une position unique sur le génome (nommés les reads mappés de façon unique) et avec un minimum de mismatches permis (exemple, jusqu'à deux mismatches), sont gardés pour les analyses en aval. Dans le cas où un read de ChIP-Seq est aligné à des positions multiples sur le génome, la solution générale est de lui affecter, de façon aléatoire, une des positions. Souvent, le ratio du nombre des reads alignés de façon unique sur le nombre total des reads ChIP-Seq peut être une évaluation de la qualité de la librairie. À partir des statistiques des ensembles de données ChIP-Seq disponibles au publique, des ratios de plus de 50% suggère une bonne qualité de la librairie.

3.11. Identification de Pics "peak calling"

Après l'alignement, les séquences de reads mappées au génome sont sujettes au peak calling. Ce dernier consiste à détecter les régions avec un enrichissement significatif des signaux ChIP par rapport au background (exemple, un contrôle si disponible). Ces signaux correspondent aux régions génomiques présentant un grand nombre de reads alignés avec une certaine signification statistique par rapport à un contrôle.

Pour la majorité des expériences ChIP-seq single-end, les fragments d'ADN sont séquencés à partir de leurs extrémités 5'. Comme résultat, des distributions bimodales, encerclant le vrai site de liaison, sont formées à partir des reads alignés sur les brins + et - (les courbes rouges et bleues dans la Figure 14 et 15, respectivement). Pour détecter de façon précise le vrai site de liaison, certains peak callers modélisent empiriquement la distance entre les modes des brins + et - (Zhang, 2008 ; Kharchenko 2008), et étendent les reads vers leur direction 3' afin d'atteindre la distance estimée des fragments d'ADN d'origine (voir Figure 15). Ainsi, la pile des reads étendus forme un pic (Figure 15) et son sommet représente la localisation la plus probable de la liaison. Une mesure de contrôle de la qualité à cette étape est la capacité des peak callers à modéliser correctement la distance entre le mode +/- . L'échec dans la modélisation de cette distance suggère des biais potentiels dans la sonication et les étapes de construction des librairies, ou que le facteur de transcription (ou la protéine d'intérêt) se lie à des régions diffuses plutôt que des "point sources".

Après cette étape de localisation des sites de liaison potentiels, les peak callers calculent la signification statistique (par exemple, la p- value) du niveau de l'enrichissement des signaux ChIP dans des régions sélectionnées par rapport au

modèle du background. À cause de la nature discrète des données NGS, la plupart des programmes adoptent les distributions binomiales (DA, 2008), Poisson (Zhang, 2008 ; Zang, 2009 ; Rozowsky, 2009), binomiale négative (presque équivalente à la Poisson basée sur une moyenne locale) (Ji, 2008 ; Ji, 2010) ou la modélisation basée sur la simulation (Kharchenko, 2008 ; Fejes 2008) pour calculer cette signification statistique. En principe, ces différentes distributions sont identiques. Par exemple, la binomiale négative est une Poisson généralisée, et une Poisson dynamique, basée sur λ locale empirique, est une version plus généralisée de la binomiale négative, etc. Par conséquent, elles offrent une sensibilité et une spécificité similaires. Par contre, les divers peak callers ont leur propre rationalité et diffèrent les uns des autres dans la précision de la localisation des sites de liaison prédits. En effet, ces algorithmes n'identifient pas les sommets des peaks aux mêmes positions, parce que les reads enrichis ne s'étalent pas de façon identique mais en fonction de la méthode de distribution utilisée pour définir ces régions.

Les peak callers largement utilisés comprennent MACS (Zhang, 2008) , CisGenome (Liu, 2011), QuEST (Valouev, 2008), FindPeaks (Fejes, 2008), PICS (Zhang, 2010).

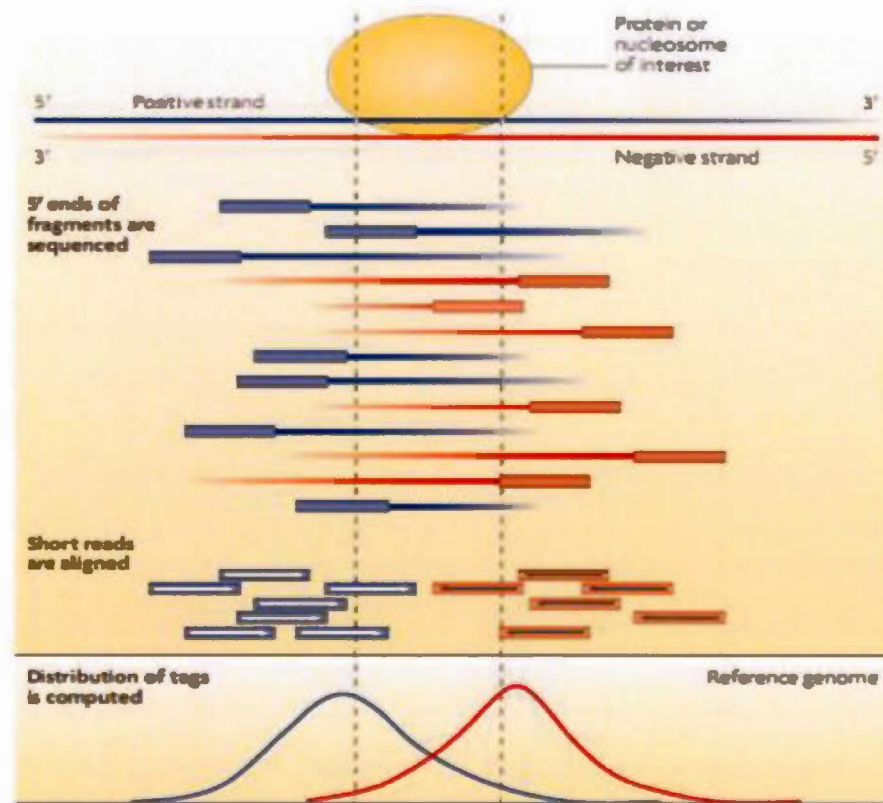


Figure 14. Profil des fragments d'ADN issus d'une expérience ChIP séquencés à partir de leur extrémité 5'. L'alignement de ces tags (appelés aussi des reads) sur le génome de référence produit deux pics (un sur chaque brin) qui flanquent l'emplacement de la liaison de la protéine ou du nucléosome d'intérêt. (Park, 2009)

3.11.1. Recherche des régions enrichies

La recherche des régions enrichies consiste à identifier les régions où les séquences reads étendues se chevauchent (Fejes ,2008) (Rozowsky, 2009) (Barski, 2007), ou alors où elles sont trouvées dans une certaine distance d'un clustering fixe (Johnson, 2007), (Kallin, 2009 ; Tuteja, 2009 ; Mortazavi, 2008). Une autre méthode,

couramment utilisée, se base sur un simple décompte des lectures dans une fenêtre d'une largeur déterminée à travers le génome, et selon un certain enrichissement par rapport au contrôle. Cette approche est connue sous le nom d'algorithme de glissement de fenêtre (Zhang, 2008 ; Da, 2008 ; Chen, 2008 ; Blahnik, 2009 ; Ji, 2008 ; Qin 2010 ; Spyrou, 2009; Kharchenko, 2008 ; Jothi,2008). Comme cet estimateur de densité, de type histogramme, peut produire des effets de bordure qui dépendent de la fenêtre ou de la taille de la boîte, certains programmes utilisent à la place un estimateur de la densité du noyau gaussien "Gaussian kernel density estimator " (G-KDE). Celui-ci génère une estimation continue de la couverture (Valouev, 2008 ; Boyle, 2008 ; Lun, 2009). Toutes ces méthodes précisent certains critères de hauteur minimale, à laquelle l'enrichissement est considéré comme important, et certains espacements minimaux, où des fenêtres adjacentes des clusters ou maxima locaux (G-KDE) sont fusionnées en une seule région de pic.

3.11.2. Méthode de la construction des profils

La méthode des algorithmes peak callers consiste à construire un profil de distribution des lectures sur chaque brin, puis à combiner ces deux profils en un seul, déterminant ainsi la région de liaison du facteur (Boyle et al, 2008 ; Valouev et al., 2008) (voir figure 15). Pour construire ce profil, la plupart des programmes effectuent un ajustement des séquences reads afin de mieux représenter le fragment d'ADN d'origine, soit en déplaçant ("shifting") les tags dans la direction 3' (Zhang, 2008 ; Valouev, 2008 ; DA, 2008) ou par l'extension des reads à la longueur estimée des fragments originaux (Johnson, 2007, Robertson, 2007; Chen, 2008 ; Fejes, 2008 ; Kallin,2009 ; Rozowsky, 2009 ; Tuteja, 2009 ; Blahnik, 2009). Cette dernière approche est normalement la plus précise, mais elle implique une estimation de la taille d'origine des fragments séquencés, tout en faisant l'hypothèse que cette taille soit uniforme. Par conséquent, lorsque la longueur moyenne des fragments peut être

déduite avec précision (soit par calcul ou empiriquement), la densité combinée forme un seul pic, où, le sommet correspond étroitement au site de liaison.

Si, ce sont les technologies de séquençage paired-end qui ont été utilisées, la longueur du fragment peut être mesurée directement et de façon juste permettant ainsi de déterminer plus précisément les sites de liaison. Cette fonction est actuellement soutenue par seulement une poignée d'algorithmes peak calling. Parmi ces derniers, il y a ceux qui supportent aussi bien les single-end que les paired-end (par exemple, MACS (Zhang, 2008)), et ceux qui sont spécifiquement conçus pour améliorer la sensibilité et la spécificité dans le séquençage paired-end (par exemple, SIPeS (Wang, 2010)).

Les algorithmes actuels ignorent les faux-positifs comme les agrégations ou empilements de reads, toutefois, afin d'améliorer la spécificité, les lectures dupliquées (même extrémité 5') peuvent être retirées avant le peak calling.

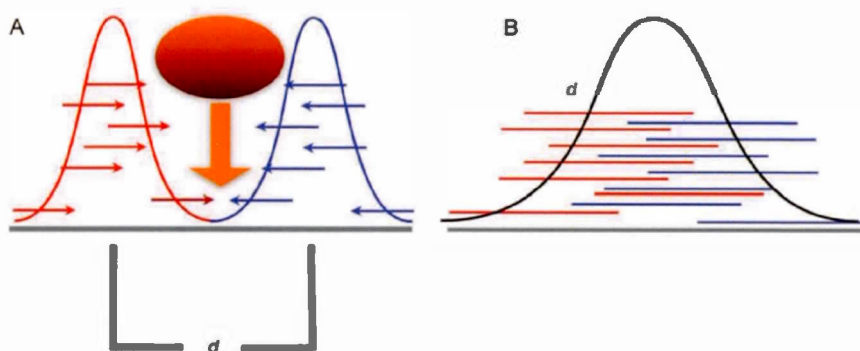


Figure 15 : Schéma de la modélisation des pics pour une expérience Chip-Seq. Les distributions bimodales des séquences de reads + et - (les flèches rouges et bleues, respectivement) encerclant un centre de liaison d'un facteur de transcription (marqué par la flèche jaune verticale). La distance (d) entre les sommets des distributions + et - est considérée comme une estimation de la longueur des fragments d'ADN précipités par l'anticorps. (B) Le signal d'enrichissement ChIP peut être obtenu en calculant le nombre des séquences de reads + et - étendues à la distance estimée (d) à chaque base (Shin, 2013).

3.11.3. FDR et autres mesures de qualité

Peu importe le logiciel utilisé, les pics détectés doivent être vérifiés en termes de la qualité. Pour ce faire, le taux de fausse découverte " false discovery rate " (FDR) ou le " fold change " par rapport au background (exemple, le nombre de reads contrôle dans la même région), sont souvent utilisés.

Le FDR est défini comme étant la proportion de pics faux positifs attendue dans une liste de pics détectés (Benjamini, 1995; Storey, 2002; Storey, 2003). Plusieurs peak callers fournissent un FDR empirique (Zhang, 2008; DA, 2008; Fejes, 2008; Valouev, 2008; Tuteja, 2009; Johnson, 2007), un FDR de modèle estimé (Ji, 2008; Ji, 2001; Zang, 2009; Rozowsky, 2009) ou alors une q-value (FDR minimum à un cut-off donné de p-value) pour chaque pic.

Le FDR empirique peut être calculé par le nombre de pics contrôles passant un certain cut-off divisé par le nombre de pics ChIP-seq passant le même cut-off. Par contre, le FDR de modèle estimé est calculé par permutation ou échantillonnage aléatoire. Généralement, un FDR de 5% est la valeur la plus communément acceptée pour les pics de bonne qualité.

De son côté, le fold change est aussi une mesure intuitive de la qualité des pics. Il représente le ratio du nombre de reads, dans la région du pic, entre l'échantillon ChIP-Seq et l'échantillon contrôle. Un fold change de 5, comme un cut-off raisonnable, est généralement recommandé et un nombre assez de pics (exemple, >50%) avec plus de 20-fold est un indicateur d'un bon enrichissement ChIP.

Les peak callers existants ont de nombreux paramètres réglables par l'utilisateur qui peuvent grandement influencer sur le nombre et la qualité des pics appelés. La p-value ou le FDR pourraient être extrêmement touchés par le modèle statistique utilisé, la profondeur de séquençage, ou le nombre précis de sites de liaison dans le génome.

Ainsi, l'utilisation de la même valeur p (p -value) ou du même seuil FDR n'assure pas que les nombres de pics appelés soient comparables entre les librairies et entre les différents peak callers (Szalkowski, 2011). Ainsi, la meilleure approche est le seuil du taux de la découverte reproductible (IDR) (Li, 2011), qui, avec l'analyse de motif, peut aussi aider à choisir le meilleur algorithme peak-calling et le réglage des paramètres.

Une description plus détaillée et une comparaison entre différents peak callers sont présentées dans les Tableau 8 et les Figures 16 et 17. Pour plus d'informations, le lecteur pourrait se référer aux études séparées (Park, 2009 ; Garber, 2011; Pepke, 2009 et Wilbanks, 2010).

Table 8 : Exemples de "peak callers" employés dans ChIP-seq (Bailey, 2013)

Logiciel	Version	Disponibilité	Régions pointes (pics)	Régions étendues (domaines)
BayesPeak (Spyrou, 2009)	1.10.0	http://bioconductor.org/packages/release/bioc/html/BayesPeak.html	oui	
BEADS [§] (Cheung, 2011)	1.1	http://beads.sourceforge.net/	oui	oui
CCAT (Xu, 2010)	3.0	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperCCAT.htm		oui
CisGenome (Hongkai, 2011)	2.0	http://www.biostat.jhsph.edu/~hji/cisgenome/	oui	
CSAR (Muino, 2011)	1.10.0	http://bioconductor.org/packages/release/bioc/html/CSAR.html	oui	
dPeak (Chung, 2013)	0.9.9	http://www.stat.wisc.edu/~chungdon/dpeak/	oui	
GPS/GEM (Guo, 210)	1.3	http://cgs.csail.mit.edu/gps/	oui	
HPeak (Qin, 2010)	2.1	http://www.sph.umich.edu/csg/qin/HPeak/	oui	
MACS (Zhang, 2008)	2.0.10	https://github.com/taoliu/MACS/	oui	oui
NarrowPeaks (Mateos, 2015)	1.4.0	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	oui	
PeakAnalyzer/ PeakSplitter [§] (Salmon-Divon, 2010)	1.4	http://www.bioinformatics.org/peakanalyzer	oui	
PeakRanger (Feng, 2011)	1.16	http://ranger.sourceforge.net/	oui	oui

PeakSeq (Rosowsky, 2009)	1.1	http://info.gersteinlab.org/PeakSeq	oui	
polyaPeak (Wu, 2014)	0.1	http://web1.sph.emory.edu/users/hwu30/polyaPeak.html	oui	
RSEG (Lampy, 2011)	0.6	http://smithlab.usc.edu/histone/rseg/		oui
SICER (Xu, 2014)	1.1	http://home.gwu.edu/~wpeng/Software.htm		oui
SIPeS (Wang, 2010)	2.0	http://gmdd.shgmo.org/Computational-Biology/ChIP-Seq/download/SIPeS	oui	
SISSRs (Jothi, 2008)	1.4	http://sisrs.rajajothi.com/	oui	
SPP (Kharchenko, 2008)	1.1	http://compbio.med.harvard.edu/Supplements/ChIP-seq/	oui	oui
USeq (DA, 2008)	8.5.1	http://sourceforge.net/projects/useq/	oui	
ZINBA (Rachid, 2011)	2.02.03	http://code.google.com/p/zinba/	oui	oui

[§] uniquement pour le post-traitement.

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific density	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromosome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromosome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
			Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data				

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method excludes putative duplicated regions, no treatment of deletions

Figure 16: Programmes de "peak calling" sélectionnés pour évaluation.

Des Programmes open-source capables d'utiliser des données contrôles ont été sélectionnés pour des tests basés sur la diversité de leurs approches algorithmiques et de leur convivialité générale. Les caractéristiques communes aux différents algorithmes sont résumées et regroupées par leur rôle dans la procédure d'appel de pics "peak calling" (les blocs colorés). Les programmes sont classés par les caractéristiques qu'ils utilisent (Xs) pour appeler des pics dans les données ChIP-Seq. Comme les listes de caractéristiques peuvent changer avec les mises à jour des programmes, La version évaluée dans cette analyse est indiquée pour chacun d'eux (willibanks, 2010).

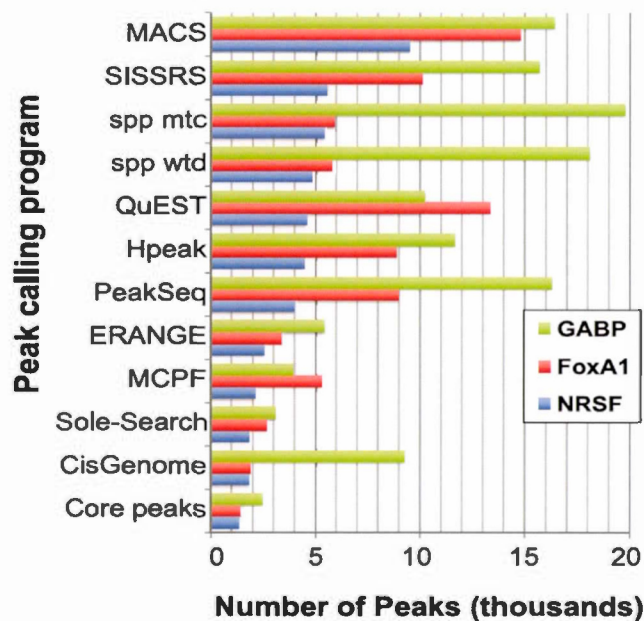


Figure 17: La quantité de pics identifiés par différents peak callers.

Lorsque lancés avec leurs paramètres recommandés ou par défaut et sur le même ensemble de données, les programmes rapportent des nombres de pics différents. Le nombre de pics rapporté est pour les ensembles de données de GABP (les barres vertes), FoxA1 (les barres rouges) et NRSF (les barres bleues). Seuls les pics identifiés par les 11 méthodes ont été calculés (les pics de base) (willibanks, 2010)

3.12. L'analyse des séquences

Elle comporte deux étapes: la recherche des motifs surreprésentés et leur identification.

3.12.1. Recherche des motifs surreprésentés

Les "peak callers" donnent en sortie plusieurs séquences contenant le motif de liaison du FT d'intérêt. Pour pouvoir les identifier, il est indispensable de faire appel à

d'autres outils, dits outils de recherche de motifs tels que MEME (Bailey, 2006), CEAS (Xuwo, 2006), Cis-Finder (Sharov, 2009), GADDEM (Li, 2009), Weeder (Pavesi, 2007), et FlexModule (Thomson, 2003) (pour plus d'outils, voir le Tableau 9). Cependant, comme le site de liaison du facteur de transcription n'est pas unique, il ne s'agit pas de détecter une séquence précise mais plutôt un patron de celle-ci. Ainsi, des modèles, comme les séquences consensus (Day, 1992) ou les expressions régulières (Myers, 1989), ont été appliqués pour la recherche et la découverte de motifs. Il est aussi possible d'utiliser une matrice de poids-position, autrement dit une PWM, un modèle plus général, présenté dans un travail fondamental antérieur (Berg, 1987 ; Von Hippel, 1988). Une PWM indique la préférence nucléotidique du facteur à chaque position du motif (voir Figure 18) (voir aussi la section analyse bio-informatique des séquences cis-régulatrices du chapitre II).

De façon général, un alignement multiple local avec peu de gap des séquences des sites de fixation de facteurs de transcription (TFBS) peut être converti en une matrice de poids position (PWM), avec les nucléotides dans les lignes et les positions des sites de liaison dans les colonnes (la longueur de l'alignement est déterminée par le nombre de colonnes, voir figure 18). Les valeurs dans la matrice montrent les préférences pour les nucléotides correspondants dans chaque position particulière de l'alignement. Dépendamment de la stratégie particulière de la construction de la PWM, la matrice peut contenir les probabilités des nucléotides ou les poids, comme des transformations log-odds des fréquences de nucléotides (stormo, 2000) (Pour plus d'information se référer à la section analyse bio-informatique des séquences cis-régulatrices du chapitre II).

Un segment d'ADN d'une longueur fixe, un mot ADN, peut alors être utilisé pour sélectionner une seule valeur réelle correspondant à chaque nucléotide particulier à chaque colonne. La somme des valeurs de la matrice sélectionnées pour le mot ADN est nommé " le score du mot " ou " le score du site de liaison ". Ce score représente la qualité du TFBS, comme reconnue par la matrice PWM, et est lié à l'affinité de liaison du FT (Berg et Von Hipell, 1987). La longueur de l'alignement, mesurée

comme le nombre de colonne dans la matrice PWM, est nommée " la longueur du motif " ou " la largeur du motif ".

Plusieurs autres modèles, ayant un plus grand nombre de paramètres que la PWM, ont été proposés (Oshchepkov et al, 2004; Levitsky et al, 2007) avec leur force de prédiction évaluée. Toutefois, les matrices de poids-position continuent d'être le modèle le plus largement utilisé pour les courts patrons de séquences TFBS des analyses des données ChIP-Seq.

Les logiciels de recherche de motifs existants diffèrent les uns des autres par leur initiation des PWMs en début de recherche. Par exemple MEME, qui utilise un algorithme EM (Expectation-Maximization) pour prédire les PWMs, commence son analyse en utilisant toutes les sous-séquences possibles, contenues dans l'ensemble de données comme PWMs, et ne garde finalement que celles qui possèdent la plus grande vraisemblance d'apparition. Bien qu'il utilise lui aussi un algorithme EM, Gadem travaille avec des dyades espacés, c'est à dire deux motifs séparés par un espace variable. De son côté, FlexModule, implanté dans CisGenome, utilise une méthode très répandue, l'échantillonneur de Gibbs (Gibbs Motifs Sampler). Concernant, CisFinder, ce dernier recherche de courtes séquences d'ADN surreprésentées et les convertie en PWM. Par la suite, il essaye de les étirer en insérant des trous et en étendant les extrémités. Weeder, quant à lui, recherche de façon exhaustive, en se basant sur les séquences consensus, des k-mers allant de 6 à 12 pb avec possibilité de plusieurs mutations (de 1 à 4) et retourne les meilleures occurrences pour chacune de ces longueurs.

Tous ces logiciels calculent une E-value pour chacune des PWMs trouvées, qui représente sa probabilité d'apparition par hasard dans les séquences.

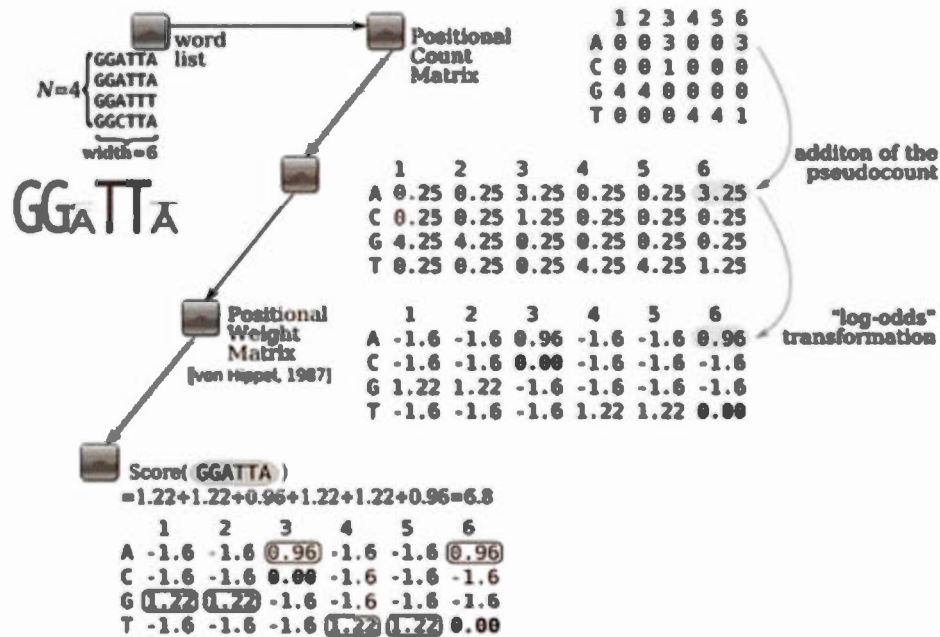


Figure 18: PWM, construite à partir d'un alignement multiple local avec peu de gap, comme le modèle basique de TFBS. La représentation du motif Logo (montré sous l'alignement) est communément utilisée pour afficher l'importance des colonnes (la hauteur globale des colonnes) et les nucléotides particuliers (la hauteur des lettres), voir (Crooks, 2004). La pseudo valeur de calcul est ajoutée pour modéliser les nucléotides fonctionnels possibles du TFBS qui manquent dans l'alignement. La transformation Log-odds (Stormo, 2000 ; Lifanov, 2003) est appliquée pour rendre les scores des colonnes individuelles additives (Kulakovskiy, 2014).

3.12.2. Identification des motifs

Une fois identifiées, les PWMs trouvées seront comparées à une base de données de sites de liaison de facteur de transcription telle que JASPAR (Sandelin, 2004), UniPROBE (Robasky, 2011) ou TRANSFAC (Matys, 2003).

Malheureusement, contrairement aux autres étapes d'analyse, il n'existe que peu d'outils pour l'identification de motifs. Les plus populaires sont TOMTOM (Gupta, 2007) et STAMP (Mahony, 2007). Ces derniers identifient les motifs en les comparant à chacune des PWMs de la base de données et un score, sous forme d'une E-value, est calculé pour chaque alignement. Plus la E-value est faible, plus la confiance accordée à l'alignement est importante. Seuls les 5 meilleurs alignements pour chacun des motifs seront par la suite envoyés à l'utilisateur, sous forme de page HTML. Bien qu'ils soient identiques du point de vue mode de fonctionnement, c'est le logiciel STAMP qui a gagné en popularité. Ceci est dû au fait que TOMTOM ne peut identifier qu'un seul motif à la fois mais aussi au fait que STAMP se distingue par sa rapidité, son avantage de présenter plusieurs options d'alignement, la possibilité de télécharger les résultats au format PDF mais aussi le grand nombre de base de données qu'il propose.

Dans notre étude nous avons utilisé un nouveau outil de recherche et d'identification de motif, le logiciel HOMER. De ce fait, nous nous sommes concentrés sur cet outil en donnant une description beaucoup plus détaillée que celles présentées pour les autres programmes sus-cités. Toutefois, pour plus de détails, le lecteur pourrait consulter les références qui leurs sont associées.

3.12.3. HOMER pour la recherche et l'identification de motifs

3.12.3.1. Introduction

HOMER est une collection d'outils qui sont communément nécessaires pour l'analyse des profils de l'expression de gène (microarray) et les expériences de l'analyse des

localisations de protéines à l'échelle du génome (ChIP-Seq ou ChIP-Chip). Initialement, HOMER a été désigné que pour la découverte de motif, mais des fonctionnalités additionnelles ont été ajoutées comme les analyses Gene Ontology et les analyses ChIP-Seq.

Ici, nous ne présenterons que les fonctionnalités pour l'analyse des motifs. Les informations décrites sont extraites du site HOMER: <http://homer.salk.edu/homer/motif/>

3.12.3.2. L'analyse de motif de HOMER

Homer contient un nouvel algorithme de découverte de motifs qui a été conçu pour l'analyse des éléments de régulation dans les applications de la génomique (uniquement l'ADN). C'est est un algorithme de découverte de motifs différentiels, ce qui signifie qu'il prend deux ensembles de séquences et essaye d'identifier les éléments de régulation qui sont spécifiquement enrichis dans un ensemble par rapport à l'autre. Pour déterminer cet enrichissement, il utilise le score ZOOPS (zéro ou une occurrence par séquence) couplé avec les calculs d'enrichissement hypergéométriques (ou binomiaux). Homer essaie également de tenir compte du biais séquencé dans le jeu de données. Il a été conçu pour les analyses Chip-Seq et celles du promoteur, mais il peut être appliqué à peu près à tout problème de recherche de motifs d'acides nucléiques.

Il y a plusieurs façons d'effectuer une analyse de motif avec Homer. En bref, il contient deux outils, `findMotifs.pl` et `findMotifsGenome.pl`, qui gèrent toutes les étapes de la découverte de motifs dans les régions promotrices et génomiques, respectivement. Ces scripts tentent de faciliter l'analyse de motifs enrichis dans une liste de gènes ou de positions génomiques (les pics ChIP-Seq). Cependant, si les fichiers de séquences à analyser sont des fichiers FASTA, `findMotifs.pl` (et `homer2`) peuvent les traiter directement.

3.12.3.3. Les étapes de l'analyse

- L'extraction des séquences (findMotifs.pl/findMotifsGenome.pl)

Si ce sont les régions génomiques qui sont fournies en entrée, l'ADN génomique approprié sera extrait. Sinon, si ce sont les numéros d'accension des gènes qui sont fournis, ce sont les régions promotrices appropriées qui seront sélectionnées.

- La sélection du Background (findMotifs.pl/findMotifsGenome.pl)

Si les séquences du background n'ont pas été explicitement définies, HOMER les sélectionnera automatiquement. Si ce sont les positions génomiques qui sont utilisées, les séquences seront sélectionnées aléatoirement à partir du génome et matchées pour la teneur en GC% (pour rendre la normalisation GC plus facile dans l'étape suivante). Par contre, si c'est une analyse basée sur le promoteur, tous les promoteurs (sauf ceux choisis pour l'analyse) seront utilisés comme background. Des backgrounds personnalisés peuvent aussi être spécifiés avec l'option "-bg <file>".

- Les séquences de normalisation GC (findMotifs.pl/findMotifsGenome.pl)

Les ensembles cibles et les ensembles background sont regroupés en fonction de leur teneur en GC (intervalles de 5%). Les séquences du Background sont pondérées pour ressembler à la même distribution de la teneur en GC observée dans les séquences cibles. Cela permet à HOMER d'éviter de trouver que des motifs riches en GC lors de l'analyse des séquences des îlots CpG. Pour effectuer une normalisation CpG% au lieu de la normalisation GC% (G + C), l'option "-cpg" est utilisée.

- L' auto-normalisation (homer2/findMotifs.pl/findMotifsGenome.pl)

HOMER offre l'autonormalisation (disponible dans la version 3 de HOMER) comme une technique pour éliminer (ou partiellement supprimer) les déséquilibres dans les séquences de courts oligo (c'est à dire, AA) en attribuant des poids aux séquences du background. Par ce moyen, la procédure tente de minimiser la différence dans la fréquence des courts oligos entre les ensembles de données cibles et ceux du background. L'auto-normalisation calcule les poids désirés pour chaque séquence du background pour aider à minimiser l'erreur. En raison de la complexité du problème, Homer utilise une approche simple, le "hill-climbing", en faisant un petit ajustement du poids du background à la fois. Elle pénalise aussi les grands changements dans le poids du background pour éviter les solutions triviales comme une augmentation ou une diminution du poids des séquences aberrantes "les outlier" aux valeurs extrêmes. La longueur des oligos courts est contrôlé par l' option "-nlen <#> "

3.12.3.4. La découverte de motifs de novo (homer2)

- Analyse des séquences input dans une table d'oligo

Le tableau d'oligo contient chaque oligo unique dans l'ensemble de données, en retenant combien de fois il se produit dans les séquences cibles et dans celles du background. Ceci est fait pour rendre la recherche des motifs (qui sont essentiellement des collections d'oligos) beaucoup plus efficace. Toutefois, ceci détruit la relation entre les oligos individuels et leurs séquences d'origine.

- L'auto-normalisation des Oligos (optionnelle)

Le concept de l'auto-normalisation est aussi appliqué à la table d'oligos. L'idée est toujours d'égaliser les oligos les plus petits (c'est à dire, 1, 2, 3 pb) dans les oligos de motifs plus longs (c'est à dire 10, 12, 14 pb etc.). Ceci est un peu plus risqué, car le nombre total d'oligos des motifs plus longs peut être très important (c'est à dire 500k

pour 10 pb, beaucoup plus pour les plus longs motifs), ce qui signifie qu'il y aura d'avantage de poids à "ajuster". Toutefois, cela pourrait aider s' il ya un biais extrême de la séquence qui est difficile à épurer sur l'ensemble de données (l'option " -olen <#>").

Après la création (et éventuellement la normalisation) du tableau Oligo, HOMER conduit une recherche globale pour "oligos" enrichis. L'idée de base dit que si un "Motif" est enrichi, alors les oligos, considérés comme faisant partie du motif, devraient également être enrichis. Tout d'abord, HOMER décèle chaque oligo possible pour l'enrichissement. Pour augmenter la sensibilité, HOMER permet des mismatches dans l'oligo lors de la recherche pour l'enrichissement. Pour accélérer ce processus, qui peut être très consommateur des ressources pour les oligos longs avec un grand nombre de mismatches, HOMER sautera des oligos si par exemple ces derniers permettent plusieurs mismatches, s'ils ont plus d'instances de background que d'instances de cibles, ou s'ils permettant plus de résultats de mismatches dans une valeur d'enrichissement inférieure.

Les options "- mis <#>" contrôlent comment de nombreux mismatches seront autorisés.

- Le calcul de l'enrichissement de Motif

L'enrichissement de motif est calculé en utilisant soit la distribution hypergéométrique cumulative ou la distribution binomiale cumulée. Ces deux statistiques supposent que la classification des séquences input (c'est à dire cibles par rapport au background) est indépendante de l'apparition de motifs. Les statistiques considèrent le nombre total de séquences cibles, de séquences du background et le nombre de chaque type contenant le motif qui est en cours de vérification pour l'enrichissement. À partir de ces chiffres, nous pouvons calculer, si nous supposons qu'il n'y a pas de relation entre les séquences cibles et le motif, la probabilité

d'observer un nombre donné (ou plus) de séquences cibles avec le motif par hasard. Les distributions hypergéométriques et binomiales sont similaires, sauf que l'hypergéométrie suppose un échantillonnage sans remplacement, tandis que la binômiale suppose un échantillonnage avec remplacement.

Le problème de l'enrichissement de motif est décrit avec plus de précision par l'hypergéométrie. Cependant, la binomiale présente aussi des avantages. La différence entre elles est généralement mineure. Toutefois, s'il ya un grand nombre de séquences et que les séquences du background \gg les séquences cibles, la binomiale est préférable car elle est plus rapide à calculer. Par conséquent, elle est la statistique par défaut pour `findMotifsGenome.pl`, où, le nombre de séquences est généralement plus élevé. Par contre, si un utilisateur a son propre background avec un nombre limité de séquences, il pourrait être une bonne idée de basculer sur l'hypergéométrie ("-h" est l'option pour forcer l'utilisation de l'hypergéométrie). D'ailleurs, comme `findMotifs.pl` s'attend à un plus petit nombre pour l'analyse de promoteur, il utilise l'hypergéométrie par défaut.

Il est aussi important de noter que comme Homer utilise un tableau d'Oligo pour beaucoup de calculs internes de l'enrichissement de motif, où, il ne sait pas explicitement combien de séquences originales contiennent le motif, il se rapproche de ce nombre en utilisant le nombre total d'occurrences de motifs observés dans les séquences cibles et du background. Il suppose que les occurrences sont réparties de façon égale parmi les séquences cibles ou celles du background (quelques-unes des séquences sont susceptibles d'avoir plus d'une occurrence). Il utilise le nombre attendu de séquences pour calculer la statistique de l'enrichissement (la sortie finale reflète l'enrichissement réelle basée sur les séquences originales).

– Optimisation de la matrice

HOMER prend la majorité des oligos enrichis de l'étape d'optimisation globale, les transforme en simples matrices de probabilité de position spécifique, et les optimise avec un algorithme d'optimisation sensible locale. Cette étape est effectuée pour

chaque oligo séparément. Elle créera la "matrice de probabilité de motif" et déterminera aussi le seuil de détection optimal pour maximiser l'enrichissement du motif dans les séquences cibles vs les séquences du background. Le seuil de détection se fait simplement en marquant chaque oligo dans les données de la matrice de probabilité, puis en triant les oligos par leur similitude avec la matrice. HOMER quitte alors la liste, ce qui diminue efficacement le seuil de détection en incluant de plus en plus d'oligos jusqu'à ce qu'un enrichissement optimal soit trouvé. Après cette étape, HOMER créera plusieurs nouvelles matrices de probabilités basées sur les oligos trouvés dans différents seuils de détection et vérifiera lequel a le plus grand enrichissement. Ce processus est répété jusqu'à ce que l'enrichissement ne puisse plus être amélioré, produisant un motif final.

- Masquer et répéter

Après que le premier "oligo potentiel" soit optimisé dans un motif, les séquences liées par le motif sont retirées de l'analyse et l'oligo potentiel suivant est optimisé pour le deuxième motif, et ainsi de suite. Cette opération est répétée jusqu'à ce que le nombre désiré de motifs soit atteint ("-S <#>", par défaut: 25).

C'est là que réside la différence clé entre l'ancienne (homer) et la nouvelle (homer2) version de HOMER. L'ancienne version masque tout simplement les oligos liés par le motif dans la Table Oligo. Par exemple, si le motif était GAGGAW alors GAGGAA et GAGGAT seraient retirés de la table Oligo pour éviter d'avoir avec le prochain motif trouvé les mêmes séquences. Toutefois, si GAGGAW a été enrichi dans les données, il ya une bonne chance que tout oligo 6-mer comme nGAGGA ou AGGAWn serait également peu enrichi dans les données. Ceci avait le risque d'inciter HOMER à trouver plusieurs versions du même motif et de fournir un peu de confusion dans les résultats. Pour éviter ce problème dans la nouvelle version de HOMER (homer2), une fois qu'un motif est optimisé, HOMER revisite les séquences

originales et masque sur les oligos qui composent l'instance du motif ainsi que les oligos immédiatement adjacents au site qui se chevauchent avec au moins un nucléotide.

Cela permet de fournir des résultats beaucoup plus propres, et permet une plus grande sensibilité avec des motifs co-enrichis.

3.12.3.5. La recherche de l'enrichissement des motifs connus (liste homer2)

– Charger la librairie de Motif

Afin de chercher des motifs connus dans les données, HOMER charge une liste de motifs préalablement déterminés à partir des données précédentes. L'ajout de motifs est également permis en les spécifiant sur la ligne de commande ("-mknown <fichier>") ou en éditant le fichier principal ("data / knownTFs / known.motifs"). En raison de la qualité de motif (qui peut être faible) et particulièrement au fait de la nécessité d'un seuil de détection, HOMER ne filtre pas tout TRANSFAC mais il le parcourt de façon partielle et intelligente.

– Épuration de chaque Motif

Pour trouver l'enrichissement pour chaque motif, HOMER scanne chaque séquence sur les instances du motif et calcule l'enrichissement final en considérant combien de séquences cibles vs séquences du background sont considérées «liées». Le calcul ZOOPS (zéro ou une occurrence par séquence) est utilisé ainsi que l'hypergéométrie ou la binomiale pour le calcul de la signification.

3.13. Les outils de visualisation

Les données ChIP-seq peuvent être visualisées sur un navigateur de génome soit comme des profils de signaux ou alors des pics appelés. Un navigateur de génome est une application basée sur le Web, qui, en plus des fonctions standards, fournit d'autres informations génomiques importantes, incluant les pistes "tracks" pour l'annotation des gènes (par exemple, RefSeq ou gènes connus UCSC), la conservation évolutive, les SNPs annotés (Fujita, 2011 ; Karolchik, 2004), et les données à partir de "NIH funded genomics consortia" comme ENCODE (Raney, 2011).

L'environnement de visualisation le plus largement utilisé est le navigateur de génome de l'Université de Californie à Santa Cruz (UCSC) (<http://genome.ucsc.edu>) (Kent, 2002) (voir Figure 19).

Comme un serveur web, le navigateur du génome UCSC a des limitations de vitesse de réponse, qui sont principalement liées à la capacité de traitement du serveur et la connexion Internet. En revanche, les navigateurs du génome autonomes tels que IGV (<http://www.broadinstitute.org/software/igv/home>) (Robinson, 2011) et IGB (<http://www.bioviz.org/igb/>) (Nicol, 2009) permettent de visualiser de grands ensembles de données génomiques à partir d'un ordinateur local ou un entrepôt de données à distance et de naviguer rapidement à de multiples échelles. Ils peuvent aussi afficher au même temps d'autres types de pistes de données génomiques comme les reads NGS alignés, les mutations, le nombre de copies, l'expression des gènes, la méthylation de l'ADN, ainsi que les annotations de gènes (Robinson, 2011). D'autres navigateurs de génome NGS sont également disponibles et le lecteur peut se référer à leurs publications individuelles (Ji, 2008 ; Ji, 2011 ; Donlin, 2009 ; Podicheti, 2011 ; Huang, 2008 ; Milne, 2010 ; Nicol, 2009 ; Bao, 2009 ; Lewis, 2002).

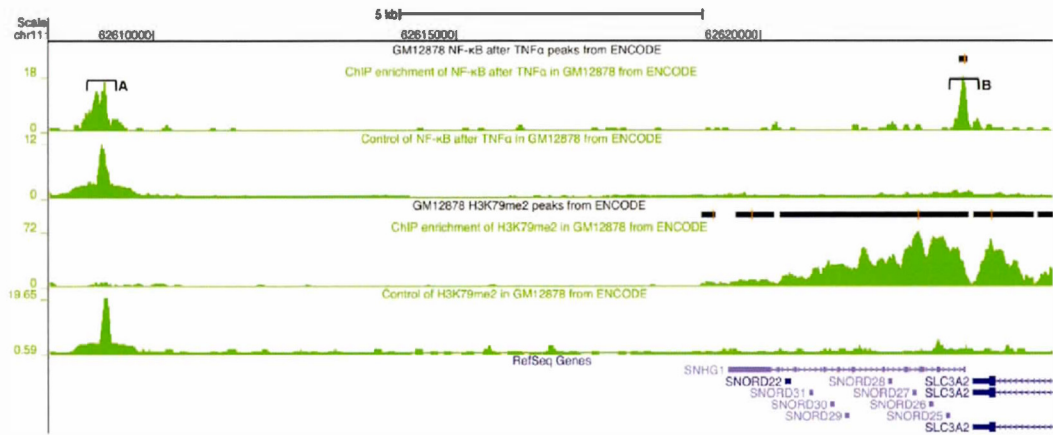


Figure 19 : Une image du navigateur de génome UCSC des enrichissements ChIP de NF-kB et de H3K79me2 dans la lignée cellulaire lymphoblastoïde GM12878 de l'humain à partir des données ENCODE. La piste du gène en bas montre les gènes RefSeq près des sites de liaison. Alors que le NF-kB montre des motifs de liaison pointus (la 1ère piste verte à partir du haut), H3K79me2 diffuse sur une large région (la 3e piste verte). Les barres noires horizontales représentent les pics appelés par MACS. Les pistes contrôles des entrées d'ADN (les 2e et 4e pistes vertes) sont utilisées pour modéliser le bruit de fond incluant le biais de la chromatine dans ChIP-Seq. Le degré de confiance de la détection de site de liaison est estimé en tenant compte du bruit de fond du contrôle. Par conséquent, la région A (près du chr11 62610000) de NF-kB n'est pas identifiée comme un pic parce que son niveau du bruit de fond est relativement élevé par rapport à la région B (près chr11 62625000) bien que les enrichissements ChIP de ces régions soient semblables (Shin, 2013).

3.14. Les Mesures de qualité

Une quantité importante d'informations génomiques sont stockées dans le domaine public tel que la conservation au cours de l'évolution, les motifs de liaison des facteurs de transcription et les annotations de gènes. Dans la section suivante, nous allons discuter de l'intérêt de l'utilisation de répliques ainsi que de la façon d'utiliser ces connaissances pour évaluer la qualité des données ChIP-seq.

3.14.1. Utilisation de répliques

Pour assurer que les résultats expérimentaux sont reproductibles, il est recommandé d'effectuer au moins deux répliques biologiques pour chaque expérience ChIP-seq et d'examiner la reproductibilité des reads et des pics identifiés (Landt, 2012 ; Rozowsky, 2009). Intuitivement, les enrichissements ChIP réussis devraient être cohérents entre les répliques biologiques en présentant des profils de signaux et des régions de pics semblables. Ainsi, les expériences répliques peuvent aider à identifier des pics avec plus de confiance.

L'une des mesures de cohérence analysée, et qui peut être facilement représentée par un diagramme de Venn (voir Figure 20), est le pourcentage (par exemple, > 50%) de pics se chevauchant entre deux répliques. La deuxième mesure concerne la reproductibilité des reads. Elle est basée sur le calcul du coefficient de corrélation de Pearson (PCC) du nombre de reads alignés à chaque position génomique (Bardet, 2011) (par exemple, plus de 0,6 entre les répliques) sur des intervalles génomiques sélectionnés (par exemple, tous les 1 kb ou 2 kb) ou sur l'union des régions de peaks des répliques. L'intervalle du PCC est typiquement de 0.3–0.4 (pour des échantillons sans corrélation -indépendants-) à 0.9 (pour des échantillons répliques dans des expériences de haute qualité). Des valeurs basses suggèrent que l'une ou les deux répliques sont de mauvaise qualité. Cependant, cette quantité peut être dominée par un petit nombre de régions hautement enrichies, dans ce cas, elle ne pourrait pas refléter la reproductibilité pour les régions qui sont moins enrichies (Łabaj, 2011). Ainsi, avant de calculer le PCC, il est important d'enlever les régions artéfacts avec de hauts signaux ChIP, tels que les régions proches des centromères, des télomères, des répétés satellites, et les régions de la liste noire de ENCODE et 1000 Genomes.

Afin de mesurer la reproductibilité au niveau du peak calling, le ENCODE / consortiums modENCODE a proposé une troisième mesure statistique de cohérence de réplicats appelée "Irreproducible Discovery Rate" (IDR) (Li, 2011). IDR mesure la proportion de pics incohérents sur les pics cohérents à un certain seuil entre deux ensembles de pics identifiés dans une paire de réplicas (Li, 2011). En plus, il analyse si les intervalles de peaks (basés sur la p-value ou FDR) sont corrélés ou non. Cette analyse a pour résultat le nombre de pics qui passent un seuil de reproductibilité spécifié par l'utilisateur (exemple, $IDR = 0.05$) (Li, 2011).

Il a été rapporté que l'utilisation d'une métrique basée sur la reproductibilité (IDR) plutôt qu'une métrique basée sur l'enrichissement (e.g., FDR or p-value) rend les nombres de pics déclarés plus comparables entre les expériences (Landt, 2012). En plus, l'analyse IDR peut aussi être utilisée pour comparer et sélectionner des pics callers (Chen, 2012 ; Li, 2011) et identifier des expériences de basse qualité (Landt, 2012).

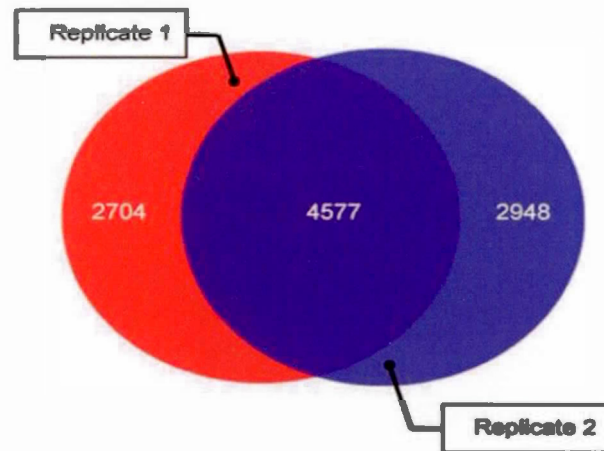


Figure 20 : Une analyse d'échantillon de deux répliques pour le facteur de transcription NF-κB d'une expérience ChIP-seq à partir de données ENCODE. Le diagramme de Venn des ensembles de pics identifiés à partir des deux répliques individuelles au même cut-off pour le peak calling (7281 and 7525 pics, respectivement). Une large intersection indique une haute cohérence entre les répliques (exemple, > 50%) (Shin, 2013).

3.14.2. La conservation au cours de l'évolution

Les éléments cis-régulateurs qui contiennent les sites de liaison de FT (TFBS) sont en général des régions hautement conservées. Par conséquent, les séquences des pics, identifiés dans une expérience ChIP-seq réussie, affichent une conservation plus élevée que les séquences du background dans le génome. Les PhastCons (Li, 2011), téléchargés à partir de UCSC genome browser, fournissent les scores précalculés de la conservation évolutive à chaque base dans un génome de référence. Ainsi, lors de l'alignement, un bon pic ChIP-seq d'un facteur de transcription présente des scores moyens de conservation des PhastCons près du sommet ou du centre plus élevés que les régions avoisinantes (voir Figure 21).

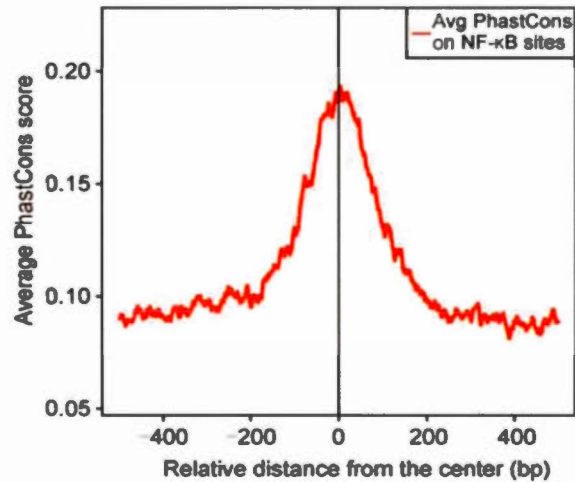


Figure 21 : Les scores PhastCons moyens au niveau des sites de liaison du TF peuvent être utilisés pour évaluer la qualité des pics de ChIP-seq. Le tracé indique que le centre des sites de liaison du facteur de transcription (NF-κB) est plus conservé du point de vue de l'évolution que le background (Shin, 2013).

3.14.3. Les motifs de liaison des FTs à l'ADN

Les facteurs de transcription se lient à l'ADN d'une manière spécifique à la séquence. Par conséquent, une expérience ChIP-seq réussie doit être caractérisée par des pics présentant un enrichissement significatif du motif de liaison du FT d'intérêt. Comme vu précédemment, les motifs de liaison peuvent être représentés comme une matrice de poids-position (PWM) ou comme une séquence logo, qui tous deux, indiquent la préférence nucléotidique du facteur à chaque position du motif.

Actuellement, plusieurs bases de données de motifs de liaison de FT connus sont accessibles au public (Robasky, 2011 ; Bryne, 2008). En outre, avec l'apparition de nouvelles technologies et de données génomiques, des groupes expérimentaux et computationnels continuent à améliorer des motifs de FT ou à caractériser d'autres restés jusqu'à lors inconnus (AlQuraishi, 2011 ; Nutiu, 2011). Des séquences motifs

enrichies peuvent être identifiées à partir de méthodes de découverte de novo ou alors en scannant les pics aux motifs connus (Bailey, 2011 ; Ma, 2012). Un motif identifié aurait une confiance élevée si ses occurrences sont localisées plus aux sommets ou aux centres des pics plutôt que dans les régions avoisinantes (Meyer, 2011). Une procédure commune pour la recherche de motif est de se concentrer sur les pics en haut du classement (typiquement le top 1000 classés selon leur indice de p-value) afin d'éviter les bruits des sites de liaison faibles.

3.14.4. Les sites DNase I hypersensibles

Il existe des régions de la chromatine qui sont sensibles plus que d'autres à une enzyme de clivage de l'ADN appelée la Désoxyribonucléase (DNase) I. L'hypersensibilité à la DNase I indique souvent une accessibilité à l'ADN associée à une diminution locale de l'occupation de nucléosomes (Bell, 2011 ; Sabo, 2004). Les sites hypersensibles à la DNase I sont largement enrichis dans les corps de gènes fortement exprimés mais aussi dans les séquences régulatrices fonctionnelles telles que les promoteurs et les enhancers, cibles de nombreux facteurs de transcription.

Récemment, le ENCODE et "Roadmap Epigenomics consortia" ont publié les données de DNase-seq de nombreuses lignées cellulaires et tissus de l'humain et de la souris (ENCODE Project Consortium, 2011 ; Bernstein, 2010). Ces données fournissent un répertoire complet des localisations de FTs, de facteurs de la chromatine, et des marques d'histones (ENCODE Project Consortium, 2011 ; Bernstein, 2010). Ainsi, les pics d'une expérience ChIP-seq réussie se chevauchent à plus de 80% avec les pics de la DNase-seq, et cela pourrait être une autre preuve de la bonne qualité des données.

3.14.5. Utilisations des régions à haute occupation

Des études ChIP-seq ont rapporté que certaines régions sont constamment enrichies dans presque toutes les expériences ChIP. Ces régions, appelées cibles à haute-occupation (HOT pour le mot anglais "high occupation targets"), ne contiennent pas le motif de liaison du facteur d'intérêt. Elles pourraient résulter des interactions protéine-protéine de facteurs inconnus avec le facteur d'intérêt ou d'artefacts expérimentaux durant le processus ChIP (Gerstein, 2010 ; Nègre, 2011). Pour ces raisons, il est conseillé de filtrer ces régions avant de procéder aux analyses en aval. Un pourcentage trop élevé de pics ChIP dans les régions HOT est souvent mauvais signe, et devrait soulever des questions quant à la qualité des données.

3.14.6. Les localisations des pics

La majorité des facteurs de transcription se lient au niveau des promoteurs que le reste du génome, par contre, les marques d'histone tels que H3K36me3 sont enrichies à travers les exons de gènes activement transcrits (voir Figure 22). Bien que ces propriétés varient largement selon le facteur de transcription ou la modification d'histone en cours d'analyse, exploiter et tester la compatibilité d'une connaissance a priori et les motifs d'enrichissement observés peuvent être utilisés pour évaluer la qualité des données ChIP-seq. Plusieurs packages de logiciels sont disponibles pour fournir des statistiques sommaires sur la distribution des pics ChIP-seq ou pour visualiser les signaux moyens d'enrichissement ChIP-seq sur les éléments cis-régulateurs annotés (Ji, 2011 ; Shin, 2009).

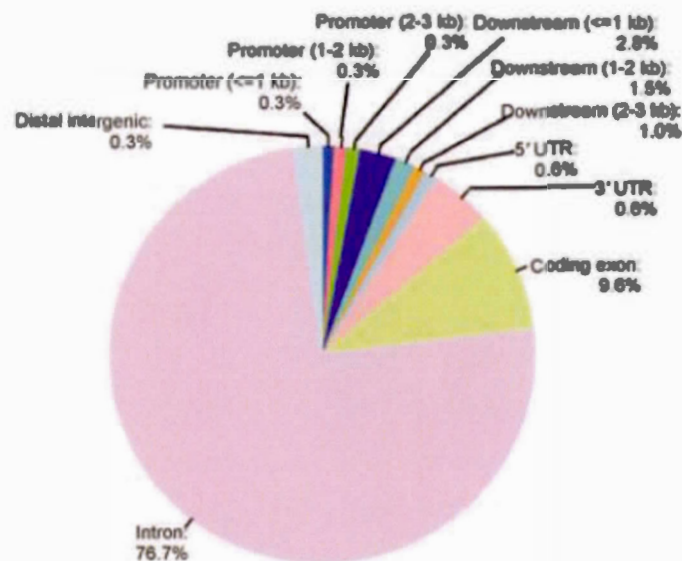


Figure 22 : Visualisation de la distribution des pics H3K36me3 sur différentes catégories d'éléments tels que les promoteurs, les UTRs, les exons codants, et les introns. Comme H3K36me3 est associé à l'élongation de la transcription, ses pics sont principalement présents dans les exons (9,6%) et les introns (76,7%) (Shin, 2013).

3.15. Analyses en Aval

En plus de trouver les localisations d'un facteur de transcription (ou d'une protéine), in vivo, et à l'échelle du génome, l'objectif d'une expérience ChIP-Seq et de tirer les informations biologiques qui y sont associées. Dans les sections suivantes, nous passerons en revue les principales analyses menées en aval de l'identification et du contrôle de la qualité des pics issus d'une expérience ChIP-Seq.

3.15.1. Annotation des pics

Le but de l'annotation est d'associer les pics de ChIP-seq à des régions génomiques fonctionnellement pertinentes, telles que les promoteurs de gènes, les sites de début de transcription (TSS), les régions intergéniques, etc. Dans une première étape, les pics et les reads, présentés dans un format approprié (par exemple un BED ou un GFF pour les pics, WIG ou bedGraph pour la couverture de reads normalisée, voir (Ji, 2011; Li, 2009)), peuvent être téléchargés dans un navigateur de génome, où les régions seront examinées manuellement à la recherche d'associations avec des caractéristiques génomiques annotées. Ainsi, si une donnée comparable, obtenue par exemple après une ChIPqPCR, est disponible, elle pourra être comparée, manuellement dans le navigateur, avec les pics et les reads ChIP-seq. Cependant, une analyse systématique peut également être effectuée en utilisant des outils dans des packages tels que BEDTools (Quinlan, 2010), CEAS (Li, 2010) ou le package Bioconductor ChIPpeakAnno (Li, 2008). Ces logiciels peuvent calculer la distance à partir de chaque pic vers le point de repère le plus proche (par exemple, TSS), ou identifier des gènes dans une distance donnée d'un pic. Les résultats de ces analyses de localisations peuvent être en corrélation avec les données d'expression. Le but, dans ce cas, est de déterminer si la proximité d'un gène à un pic est corrélée avec son expression. En outre, une analyse d'ontologie de gène, pour déterminer si le facteur ou la protéine ChIPée est impliquée dans des processus biologiques particuliers, peut aussi être effectuée en utilisant des outils comme DAVID (Lunter, 2011), GREAT (Kravitz, 2010), ou GSEA (Li, 2009). Parfois, les densités de reads par rapport à une caractéristique spécifique annotée sont tracées et comparées entre différents échantillons, révélant ainsi les différences entre des motifs de liaison de protéines (Bao, 2011).

Dans ce qui suit une présentation plus détaillée de l'annotation des pics, basée sur l'intégration d'autres données génomiques.

– La combinaison avec les profils d'expression génique

Les données ChIP-seq de FTs, des facteurs de la chromatine, et des marques d'histones ont souvent besoin d'être interprétées dans le contexte de la régulation génique, ce qui nécessite de prédire les gènes cibles qui sont régulés par le facteur et ses sites de liaison. Les méthodes pour déterminer les gènes cibles potentiels d'un facteur dépendent généralement du type de facteur et des motifs de liaison. Par exemple, la plupart des facteurs de transcription et des marques d'histone sont caractérisés par des motifs de liaison nets (courts) sur les régions proximales ou distales des sites de début de transcription (TSSs). Dans ce cas, la distance entre chaque site de liaison et son gène le plus proche tend à montrer une forte association avec le niveau d'expression ou les dynamiques dans l'expression du gène. Toutefois, bien que cette stratégie fonctionne dans de nombreux cas, elle néglige le fait qu'il pourrait y avoir des interactions à longue portée entre les éléments cis-régulateurs et leurs gènes cibles. Avec plus de compréhension sur la structure de la chromatine de haut niveau, il devrait falloir réviser la méthode de prédiction de gènes cibles. D'autre part, pour certaines autres marques d'histones avec un enrichissement omniprésent sur une région, la couverture ainsi que le niveau d'enrichissement de la marque sur le promoteur, les exons, ou le corps d'un gène sont des variables importantes pour déterminer si le gène est une cible .

L'important dans l'analyse intégrative d'une expérience ChIPseq avec des profils d'expression est de déduire si un facteur donné fonctionne principalement comme un activateur de la transcription ou comme un répresseur de celle-ci. Afin d'effectuer une telle analyse, le profil d'expression devrait être établi dans la condition où l'activité du facteur est perturbée (par exemple, par un knockdown siRNA). Si c'est une marque

d'histone qui est étudiée, l'expression génique peut être profilée de la même façon, en perturbant l'activité de l'enzyme de modification associée à la marque d'histone.

Une fois les gènes cibles de la liaison sont prédis, une autre analyse utile consiste à examiner leurs fonctions biologiques potentielles. Ainsi, l'analyse computationnelle de l'annotation et des ontologies des gènes pourrait fournir des indices importants, quant à savoir si les gènes cibles sont enrichis en processus ou en voies biologiques spécifiques.

3.15.2. La prédiction des gènes cibles

- Pour les facteurs ayant des motifs de liaison étroits (ou courts)

Le site de liaison du facteur de transcription peut être à des centaines de KB des gènes cibles, toutefois, l'effet de la liaison d'un FT sur l'expression génique s'atténue au fur et à mesure que l'on en s'éloigne (Lupien, 2008). La prédiction des gènes cibles basée sur la distance est utile pour la plupart des FTs mais aussi pour certaines marques d'histones (ayant des pics étroits sur des promoteurs ou des enhancers, par exemple régions intergéniques distal ou introniques). Les marques d'histone qui font partie de cette catégorie comprennent H3K4me3 enrichi au niveau des promoteurs, H3K4me1 enrichi au niveau des enhancers et H3K4me2 enrichi au niveau des deux. On retrouve aussi le H3R17me3, la plupart des marques d'acétylation, et le variant d'histone H2A.Z.

La prédiction des gènes cibles se base sur la recherche du site de début de transcription (TSS) le plus proche du site de liaison du FT. En général, dans le voisinage de 1 kb pour la liaison au promoteur et dans les 10 kb pour la liaison à l'enhancer. Si les profils d'expression avec une perturbation de l'activité du facteur sont disponibles, la caractérisation des gènes avec une expression différentielle significative entre les conditions de perturbation, pourrait affiner la recherche des gènes cibles. Une alternative à l'utilisation de la distance la plus proche est de

compter le nombre de sites de liaison dans un intervalle de distance donné (par exemple, 100 kb). Ainsi d'avantage de sites de liaison à proximité sont considérés comme des preuves de l'augmentation potentielle de la régulation du facteur au gène cible (Verzi, 2010). Cette approche pourrait être affinée en pondérant les différents sites de liaison par leurs distances au gène cible (Tang, 2011).

– Pour les facteurs ayant des motifs de liaison larges

Les facteurs de la chromatine et les marques d'histones associés à l'élongation ou à la repression de la transcription se lient souvent à des motifs d'enrichissement envahissants qui s'étendent sur de larges régions. Par exemple, H3K36me3 est très présent dans les exons des gènes exprimés, tandis que le complexe H3K27me3 ou PRC2 a tendance à être enrichi au niveau des gènes réprimés (The ENCODE Project Consortium, 2007). Il serait intéressant de classer les gènes en fonction de leurs niveaux d'expression et de voir l'enrichissement relatif de la marque d'histone dans chacune des classes. La Figure 23 montre la moyenne des signaux ChIP-seq de la marque d'histone H3K36me3 sur un metagène (en divisant chaque gène dans le même nombre de bins) à différents niveaux d'expression des gènes dans la lignée cellulaire lymphoblastoïde GM12878 de l'humain (Barski, 2007). Comme H3K36me3 est une marque de l'élongation de la transcription, le top 10% des gènes exprimés ont un enrichissement H3K36me3 significativement plus élevé que les 10% des gènes les moins exprimés. Ainsi, en calculant la couverture de ChIP-seq sur le corps de gène ou d'exons (normalisé par la longueur du gène ou de l'exon) et la sélection de gènes avec le top de couverture ChIP-Seq (par exemple, 30%), on peut prédire les gènes associés aux H3K36me3.

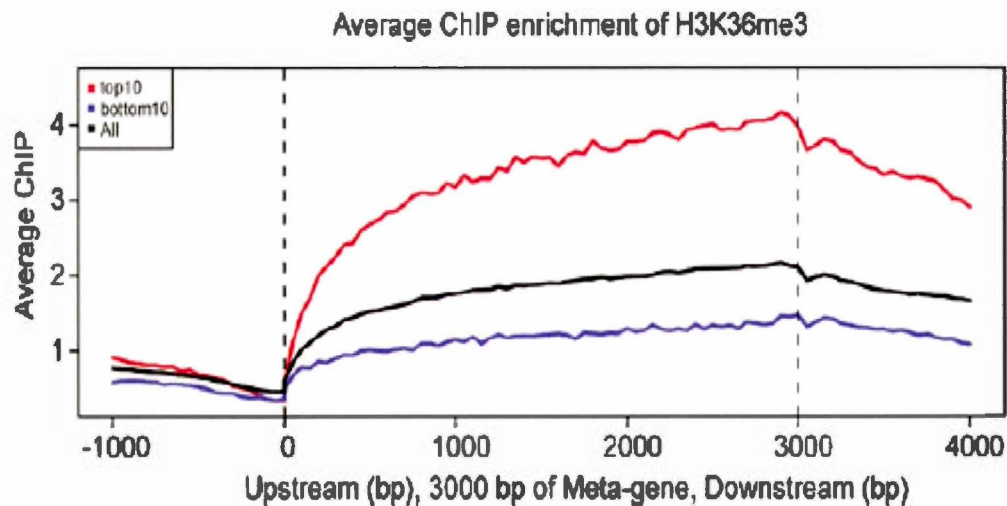


Figure 23 : La moyenne des signaux ChIP-seq de la H3K36me3 sur le metagène à différents niveaux d'expression des gènes dans la lignée cellulaire lymphoblastoïde humain GM12878. Chaque gène a été normalisé pour avoir la même longueur de 3 kb, ensuite le signal moyen ChIP a été analysé sur le méta-gène comprenant 1 kb en amont et en aval du TSS et TTS. Les lignes rouges et violettes représentent les enrichissements ChIP moyens de H3K36me3 sur les gènes hautement exprimés (top 10%, rouge) et d'expression basse (10% en bas, violet), respectivement, ce qui montre que H3K36me3 est positivement corrélé avec les niveaux d'expression de gènes (Shin, 2013).

3.15.3. La détermination du rôle du facteur

Le rôle d'un facteur, comme un répresseur ou un activateur de la transcription, peut être déduit par les changements dans l'expression de gènes cibles lorsque l'activité du facteur est perturbée par l'activation, la surexpression, le Knockdown ou le knockout. Par exemple, si les gènes ayant une expression réduite lors du knock-down ou knock-out du facteur sont plus susceptibles d'abriter des sites de liaison du facteur à proximité que les gènes avec une expression accrue, ceci sera considéré comme une preuve importante que le facteur agit comme un activateur transcriptionnel. En effet, en utilisant la distance de liaison à tous les gènes comme background, nous avons pu

voir que, dans la lignée de cellules de cancer de la prostate LNCaP, le récepteur des androgènes se lie beaucoup plus proche des gènes sur-régulés que des gènes sous-régulés, ce qui implique qu'il sert principalement d'activateur transcriptionnel. En revanche, dans la lignée cellulaire du cancer du sein MCF-7, le récepteur de l'œstrogène semble être tout aussi proche à la fois des gènes sur-régulés et sous-régulés. La signification statistique de ces observations peut être évaluée en utilisant un test non paramétrique comme le test de Kolmogorov-Smirnov d'un côté (ceci ne rentre pas dans les objectifs de notre étude).

3.15.4. L'annotation des gènes cibles basée sur "Gene Ontology" (GO) et les voies biologiques

Souvent, l'intérêt ultime d'une analyse de site de liaison des facteurs de transcription est de voir si les gènes cibles d'un TF sont enrichis pour des voies ou des processus biologiques spécifiques. De nombreux outils d'ontologies de gène ou d'analyse d'annotation sont accessibles au public et résumés sur le site Gene Ontology (<http://geneontology.org/page/go-tools-registry>). Il existe d'autres, particulièrement conviviales, qui ne sont pas dans la liste tels que DAVID (Huang, 2009; Huang, 2009), Panther (Thomas, 2003) et GREAT (McLean, 2010). Les deux premiers prennent la liste de gènes comme entrée, tandis que GREAT (McLean, 2010) prend les coordonnées des sites de liaison comme entrée et leur associe les gènes cibles en se basant sur la distance site de liaison - gènes. Malgré quelques inconvénients, cette méthode peut détecter l'enrichissement de fonctions ou de voies avec une meilleure sensibilité, car elle permet d'avoir une liste plus riche de méta-annotations de gènes (McLean, 2010). "Gene set enrichment analysis" (GSEA) effectue également une

analyse ontologique sur les gènes cibles et ses ensembles de gènes uniques fournissent des recherches d'annotation plus approfondies (Subramanian, 2005).

3.15.5. Analyse de motif

L'analyse des motifs est utile au delà du fait qu'elle identifie le motif de liaison à l'ADN dans les pics ChIP-seq du FT. D'abord, si le motif de la protéine ChIPée est déjà connu, cette analyse fournit une validation du succès de l'expérience. Et même lorsque le motif n'est pas connu à l'avance, l'identification d'un motif situé au centre d'une grande fraction de pics est indicative d'une expérience réussie. En plus, l'analyse peut identifier des motifs de liaison à l'ADN d'autres protéines, qui se lient en complexe ou en conjonction avec la protéine ChIPée. Ceci apporte plus d'indices quant aux mécanismes de régulation de la transcription. Enfin, elle est également utile avec les modifications d'histones ChIP-seq car elle peut découvrir des signaux de séquences imprévues, associées à ces marques. Le Tableau 9 et (Kuttippurathu, 2011 ; Liu, 2011) présentent une courte liste des outils disponibles au public pour l'analyse de motif.

La première étape dans l'analyse de motif est d'avoir un ensemble de séquences génomiques au format FASTA correspondant à tous les pics ChIP-seq significatifs (Goecks, 2010; Blankenberg, 2010 ; Giardine, 2005, van Helden, 2003 ; Kent, 2002). La seconde étape est celle de la découverte de motifs. Comme les algorithmes ont des forces et des faiblesses complémentaires, il est conseillé d'introduire les séquences de pics à deux ou plusieurs algorithmes de découverte des motifs (Kulakovskiy, 2010 ; Ji, 2008, Machanick, 2011 ; Thomas-Chollier, 2011). Certains de ces algorithmes font partie de pipelines qui exécutent plusieurs étapes de l'analyse de motif (par exemple, MEME-ChIP (Machanick, 2011) et le peak-motifs (Thomas-Chollier, 2011), y compris les algorithmes de découverte de motifs basés sur le mot et les

algorithmes d'enrichissement de motifs, qui peuvent identifier des motifs présents dans une petite fraction de pics. Une fois les motifs découverts, ces derniers sont comparés à ceux connus en se servant d'un logiciel de comparaison de motifs (Mahony, 2007 ; Gupta, 2007). La présence du motif du TF ChIPé est confirmée si son motif de liaison (ou sa famille de TF) est connu.

L'analyse de l'enrichissement du motif central permettra de déterminer si d'autres motifs d'ADN connus sont enrichis près des sommets des pics ChIP-seq (Bailey, 2012). Des résultats positifs indiqueraient l'existence potentielle d'autres facteurs de transcription pouvant se lier à proximité du TF ChIPé. Il pourrait également être utile d'effectuer une analyse locale de l'enrichissement de motif dans les régions centrées sur des sites génomiques, telles que les sites de début de transcription chevauchés par les pics ChIP-seq (Bailey, 2012). En outre, une analyse d'espacement de motifs pourrait révéler des distances et des dispositions préférées de paires de motifs. Ces détections peuvent être indicatives d'interactions physiques entre TFs (Whittington, 2011). Enfin, les motifs prédits peuvent être alignés et visualisés dans les emplacements génomiques des motifs dans chacune des régions de ChIP-seq (Grant, 2011, Hertz, 1999). Dans cette étape, les motifs découverts ou enrichis sont utilisés pour scanner les régions de pics ChIPseq, et les coordonnées des matchs pourront être téléchargées sur un navigateur de génome pour la visualisation.

La découverte de motifs a déjà été suggérée comme une mesure de QC importante de l'expérience ChIP-seq, et elle peut aussi être utilisée pour prédire les facteurs collaborateurs (Zhang, 2011). La collaboration ou l'association entre plusieurs TFs "Crosslinking" stabilise non seulement la liaison covalente entre TFs et l'ADN, mais aussi celle entre les facteurs de transcription en interaction. Ainsi, l'ChIP d'un seul des ces facteurs permet d'enrichir les cibles des facteurs collaborateurs. Si la recherche de motif sur les pics d'un facteur a permis de trouver non seulement le motif connu de ce

dernier, mais aussi un motif significativement enrichi d'un autre facteur, on suggère que les deux facteurs peuvent interagir. Parfois, de nombreux facteurs avec les mêmes domaines de liaison à l'ADN partagent des motifs similaires, afin d'identifier le facteur exact se liant au motif collaborateur nécessite l'analyse de données d'expression. Si un facteur est fortement exprimé dans la cellule et que son expression est corrélée avec un autre facteur ChIPé dans un ensemble d'échantillons de tissus reliés (connexes) (Carroll, 2005), alors on peut les considérer comme des partenaires collaborateurs putatifs. La littérature et des expériences d'interaction protéine-protéine pourraient aussi fournir des indications supplémentaires. Il est aussi possible d'utiliser ChIP-seq pour mieux analyser les co-facteurs candidats sélectionnés. Le diagramme de Venn (ou le diagramme Euler dans le cas de plus de deux facteurs) est également utilisé pour voir l'association potentielle ou colocalisation entre différents facteurs de liaison à l'ADN, lorsque les données de ChIP-seq de ces facteurs sont tous disponibles. Dans de nombreux cas, les régions co-occupées par des facteurs collaborateurs (c'est à dire, l'intersection dans le diagramme de Venn) sont considérées comme plus importantes pour la compréhension détaillée de l'évènement de la régulation des gènes régis par ces facteurs ensemble. Il existe des logiciels disponibles au public pour appeler ces régions colocalisées à partir de listes de pics de plusieurs facteurs au format BED de UCSC (Liu, 2011 ; Giardine, 2005 ; Quinlan, 2010). En outre, les heatmaps sont très utiles pour afficher des signaux ChIP de différents facteurs par l'union de leur sites de liaison (par exemple, de -1 kb à 1 kb du sommet ou du centre de liaison). Couplée avec les méthodes de clustering (par exemple, le clustering hiérarchique ou k-means), l'analyse de heatmap peut fournir une meilleure classification des sites de liaison que le diagramme de Venn en regroupant les sites de liaison avec des motifs d'enrichissement ChIP similaires à travers de multiples facteurs (Feng, 2011).

Table 9 : Exemples d'outils pour l'analyse de motif des pics ChIP-seq et leurs utilisations (Bailey, 2013)

Category	Software	Web Server	Obtain peak regions	Motif discovery	Motif comparison	Central motif enrichment analysis	Local motif enrichment analysis	Motif spacing analysis	Motif prediction/mapping
Obtaining sequences	Galaxy (Goecks,2010; Giardine,2005)	X	X						
	RSAT (Thomas-Chollier, 2012)	X	X						
	UCSC Genome Browser (Kent, 2002)	X	X						
Motif discovery +more	ChIPMunk (Kulakovskiy, 2010)	X		X					
	ChGenome (Ji, 2008)			X	X				
	CompleteMOTIFS (Kuttipurathu, 2011)	X		X	X				
	MEME-ChIP (Machanic, 2011)	X		X	X	X			
	peak-motifs (Thomas-Chollier, 2011)	X		X	X				
	Cistrome (Uu, 2011)	X	X	X		X	X		X
Motif comparison	STAMP (Mahony, 2007)	X			X				
	TOMTOM (Gupta, 2007)	X			X				
Motif enrichment/spacing	CentriMo (Bailey, 2012)	X				X	X		
	SpaMo (Whittington, 2011)	X						X	
Motif prediction/mapping	FIMO (Grant, 2011)	X							X
	PATSER (Hertz, 1999)	X							X

3.15.6. Intégration avec des données ChIP-seq de multiples organismes

De nombreuses expériences biomédicales sont réalisées sur des organismes modèles, il est donc intéressant d'examiner si des mécanismes découverts pour un facteur dans des organismes modèles, tels que le nématode ou la mouche des fruits, sont toujours valables dans des organismes plus complexes tels que l'humain. Cette question peut être résolue par comparaison de profils ChIP-seq de facteurs orthologues chez de multiples espèces. Plusieurs études ont montré que, bien que les sites de liaison des facteurs divergent largement entre les espèces, les gènes cibles des TFs hautement

conservés sont pour la plupart préservés entre les espèces (Odom, 2007 ; Schmidt 2010). De plus, certaines études récentes ont analysé les reads ChIP-seq qui s'alignent à des régions répétées du génome, et ont également révélé un lien significatif entre le facteur et le gène cible à travers l'évolution par "transposable element turnover " (Chung, 2011 Wang, 2007). Cependant, afin de détecter l'enrichissement dans les régions répétées avec plus de précision, davantage d'efforts sont nécessaires pour développer des algorithmes qui peuvent mieux utiliser les séquences de reads s'alignant à plusieurs endroits.

3.15.7. Combinaison avec des données épigénétiques ChIP-seq

Depuis l'apparition de ChIP-seq, plusieurs travaux épigénétiques ont été menés. Les profils de méthylation (Barski, 2007) et d'acétylation (Wang, 2008) d'histones ont d'abord été caractérisés dans les cellules lymphocytes T CD4 + humaines. D'autres efforts pour étudier les marques d'histones dans plusieurs états cellulaires et dans différents organismes ont donné une meilleure compréhension de la régulation épigénétique. Ceux-ci incluent notamment les travaux mod/ENCODE, Roadmap Epigenomics consortia ainsi que des études individuelles (Barski, 2007 ; Wang, 2008 ; Liu, 2011 ; Eaton, 2011). Par exemple, il a été démontré que de nombreuses modifications d'histones sont associées à des éléments importants tels que des promoteurs activés ou réprimés (H3K4me3 ou H3K27me3, respectivement (Bernstein, 2006)), des exons activement transcrits (par exemple, H3K36me3 (Kolasinska-Zwierz, 2009)) et des enhancers actifs (par exemple, H3K27ac (Creyghton, 2010) et H3K4me1 (Heintzman, 2007)).

En outre, les marques d'histones telles que H4K20me3 et H3K9me3 indiquent un état de répression transcriptionnelle sur de très grands domaines d'hétérochromatine. Bien que les différentes marques d'histones aient des spécificités d'anticorps différentes, plusieurs, en particulier les acétylations d'histones, ont des caractéristiques

d'enrichissement similaires et probablement des mécanismes de régulation redondants (Wang, 2008).

Plusieurs groupes ont tenté d'utiliser des algorithmes d'apprentissage machine tels que les modèles de Markov cachés pour inférer des modèles d'enrichissement combinatoires de marques d'histones et d'autres facteurs de la chromatine (Liu, 2011; Ernst, 2011; Ernst, 2010 ; Hoffman, 2012). Ces modèles combinatoires réduisent également la redondance dans les profils des marques d'histone. Ils peuvent être utilisés pour segmenter l'ensemble du génome en régions de signatures de chromatine distinctes, et pour prédire des éléments fonctionnels non identifiés précédemment dans le génome. Dans la même optique, une étude intéressante serait d'identifier les motifs de facteurs de transcription enrichis dans les régions enhanceurs putatives déduites des profils de marques d'histone. Si menée sur différents types cellulaires et intégrée à l'analyse de l'expression génique, celle-ci pourrait donner des informations importantes sur les activités du facteur de transcription spécifiques au type cellulaire (Ernst, 2011).

Des sites de liaison putatifs de facteurs de transcription peuvent porter des marques d'histones enhanceurs avant la liaison active du facteur de transcription. Ce motif de marque d'histone change souvent lors de la liaison. Il est généralement caractérisé par un déplacement du nucléosome au niveau du site de liaison, ainsi que par un meilleur positionnement et un marquage plus prononcé des nucléosomes flanquants le site. Par conséquent, l'analyse de motif, réalisée sur la dynamique des profils de marques d'histones au niveau de la résolution d'un nucléosome unique, peut potentiellement améliorer la précision de la prédiction du site de facteur de transcription (Meyer, 2011 ; He, 2010). Cette approche a récemment attiré l'attention comme une mesure efficace de présélection pour trouver des facteurs de transcription clés répondant aux stimulations environnementales (Verzi, 2010). De même, dans une autre étude, les dynamiques de l'état de la chromatine en fonction des profils des marques d'histone

ont été utilisées pour relier des enhancers avec leurs promoteurs cibles en se basant sur l'estimation de la corrélation. Ensuite, des prédictions ont été faites pour des FTs actifs ou répressifs, spécifiques au type cellulaire, par l'intégration des résultats de l'analyse de motifs avec les profils d'expression génique (Ernst, 2011).

3.15.8. Combinaison avec les variations génomiques et les données des maladies

Des variations dans la séquence d'ADN, telles que des polymorphismes d'un seul nucléotidique (SNP) sur les sites de liaison des facteurs de transcription, peuvent influencer l'affinité de liaison de ces derniers aux séquences d'ADN. Plusieurs études ont tenté d'évaluer l'impact potentiel des SNPs ou des indels spécifiques à l'allèle sur la structure de la chromatine ou des sites de liaison du TF (Kasowski, 2010; Birney, 2010). Ce qui ressort d'intéressant est que la majorité des loci associés aux maladies, découverts à partir des études d'association à l'échelle du génome (GWAS pour "genome-wide association studies"), sont situés sur des régions d'introns ou intergéniques distales. Les données ChIP-seq des facteurs de transcription ou des marques d'histones peuvent fournir des indices sur les mécanismes de ces SNPs de la maladie. Par exemple, ChIP-seq de H3K4me2 a permis d'identifier un enhancer, spécifique au tissu, dans les loci du risque de cancer (au niveau de la région 8q24 humaine), qui pourrait réguler l'expression Myc (Ahmadiyeh, 2010).

Les études de la variation du nombre de copies (CNV) (pour "Copy number variation") peuvent également être combinées avec l'analyse CHIP-seq (Pickrell, 2011). Toutefois, bien que l'étude de l'enrichissement de la liaison du TF dans les régions génomiques amplifiées puisse révéler de nouvelles voies pathologiques provenant des CNVs, elle est aussi capable de générer des appels de pics faux positifs (Pickrell, 2011). Par conséquent, des mesures doivent être prises afin de réduire ces faux positifs (par exemple, l'utilisation des informations CNVs provenant d'autres sources ainsi que l'utilisation de contrôles inputs de la chromatine bien appropriés).

3.15.9. Analyse des liaisons différentielles

Avec l'augmentation régulière des projets NGS et l'accumulation rapide des données Chip-Seq, un nombre croissant d'analyses comparatives entre des conditions ou des tissus est attendu. En effet, l'objet de nombreuses études ChIP-seq actuelles est la comparaison des profils de liaison des facteurs de transcription entre différentes conditions. La détection des changements dans les interactions ADN-protéines sous des conditions cellulaires distinctes est une étape cruciale dans la compréhension des réseaux de régulation derrière les processus biologiques tels que la différenciation cellulaire, l'activation des voies de signalisation et l'apparition de maladies (Kaikkonen, 2013; Martens, 2013). Toutefois, ces études comparatives ont besoin d'approches informatiques appropriées. Mais peu de méthodes ont été proposées. À quelques exceptions près, les méthodes de peak calling largement utilisées sont basées sur l'analyse des signaux individuels et n'offrent aucune fonctionnalité pour normaliser plusieurs expériences ChIP-seq.

Dans les sections qui suivent, nous présentons la problématique de l'analyse des liaisons différentielles d'un facteur de transcription entre différents types ou conditions cellulaires et les méthodes existantes dans ce domaine.

3.15.9.1. Problématique de l'analyse des liaisons différentielles

De façon générale, le problème du peak calling différentiel consiste à trouver des régions génomiques avec des changements dans l'intensité du signal ChIP-seq (issu de l'interaction d'une protéine avec l'ADN) entre deux ou plusieurs conditions

cellulaires. Malheureusement, la plupart des méthodes de "peak calling" ne peuvent analyser qu'un seul signal ChIP-seq à la fois et sont incapables d'effectuer un "peak calling" différentiel. Récemment, quelques approches basées sur la combinaison de ces peak callers avec des tests statistiques ont été proposées. Cependant, ces méthodes ne parviennent pas à détecter les modifications détaillées des liaisons protéine-ADN.

Initialement, le peak calling différentiel a été réalisé à partir des signaux ChIP-seq des peak calling individuels. Les pics détectés dans seulement une des conditions sont alors définis comme des pics spécifiques à des cellules (Heinz, 2010). Pour ce faire, l'approche la plus simple, dite **qualitative**, classe les pics de chaque échantillon comme étant communs ou uniques, en se basant sur l'existence ou non de chevauchement entre les pics dans les différents échantillons, respectivement (Fujiwara, 2009 ; Liu, 2010; Yu, 2009 Schmidt, 2010 ; Smagulova, 2011; Williams, 2011). Bien que cette méthode puisse identifier des relations générales entre des ensembles de signaux de différentes conditions, le simple chevauchement binaire de deux ensembles de pics (par exemple, avec BEDTools (Quinlan, 2010)) ne représente pas l'approche optimale (Bardet, 2011). Les résultats sont, en effet, très dépendants du cutoff, utilisé dans le peak calling, qui est difficile à choisir de façon totalement objective. Ainsi, les pics communs peuvent montrer une liaison différentielle entre les échantillons, du fait que de tels procédés ne sont pas en mesure de détecter les cas, où, des pics sont appelés dans deux types cellulaires mais présentant une augmentation (ou une diminution) significative du signal protéine-ADN dans l'une d'elles. Par exemple dans la Figure 24, basée sur des pics obtenus avec PeakSeq (Rozowsky, 2009), seul le signal nommé DP1 serait détecté comme spécifique à une cellule. D'un autre côté, des pics peuvent être identifiés comme étant uniques à un échantillon simplement parce qu'ils tombent en dessous d'un seuil arbitraire dans l'autre échantillon (voir Figure 25). Les différences dans les niveaux du background entre les échantillons confondent encore plus l'analyse.

Actuellement, l'approche la plus intéressante pour extraire l'information biologique maximale est **l'approche quantitative**. En bref, la comparaison quantitative des échantillons Chip-Seq propose l'analyse de la liaison différentielle entre les conditions en se basant sur les nombres totaux de reads dans les régions de pics ou sur les densités de reads, à savoir, le nombre de reads qui se chevauchent à des positions génomiques individuelles. Malheureusement, elle aussi se heurte à de nombreux défis (voir les détails dans la section suivante).

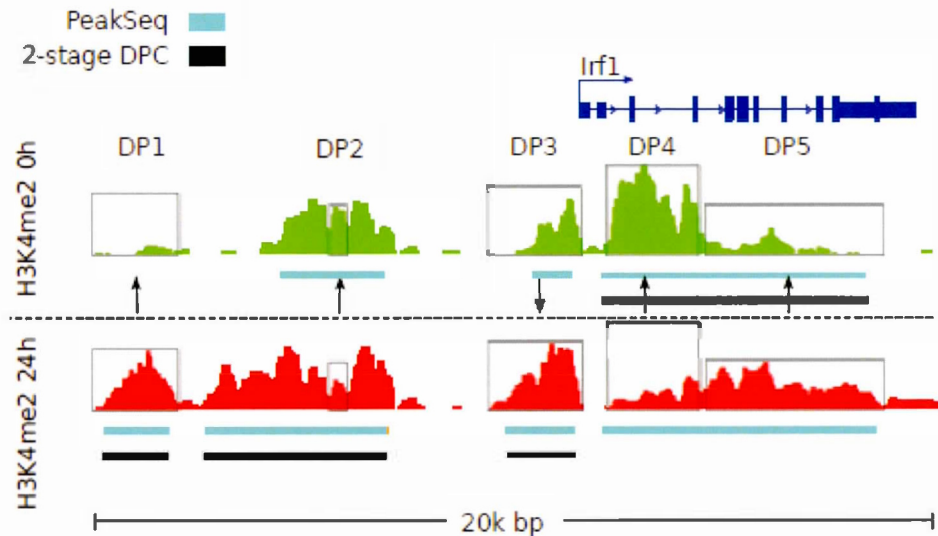


Figure 24 : Des exemples de pics différentiels putatifs après l'induction de la signalisation TLR4 (Kaikkonen, 2013).

Un exemple de deux signaux distincts ChIP-seq pour la modification d'histone H3K4me2 avant (0h, le signal vert) et 24 heures après (24h, le signal lu) l'induction de la signalisation TLR4 des macrophages autour du gène *Irf1*. Des exemples en carrés indiquent des régions, qui sont des DPs putatifs avec gain (ou la perte) du signal de ChIP-seq après 24 heures de traitement TLR4. La hauteur des carrés indique la taille du signal le plus élevé de ChIP-seq pour un DP. Les résultats de la SPC PeakSeq (barres bleues) et ceux d'un peak caller à deux étapes basé sur l'application de DESeq sur les pics de PeakSeq (barres noires) sont aussi affichés. PeakSeq détecte avec succès des pics larges décrivant le signal de ChIP-seq pour chaque cellule. Le peak caller à deux étapes peut détecter DP1 et DP3, mais ne peut pas détecter les changements dans les pics larges candidats comme DP2 ou des changements complexes dans le signal autour du corps du gène *IRF1* (DP4 et DP5). (Kaikkonen, 2013)

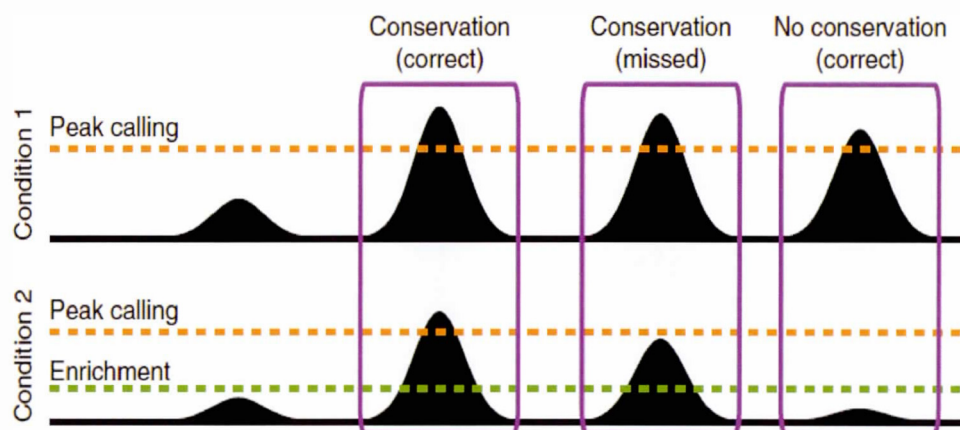


Figure 25 : Choix des seuils sensibles et la comparaison des échantillons ChIP-seq (Bardet, 2011). Au cours du "peak calling" à l'échelle du génome, seuls les meilleurs pics passent les seuils stricts requis pour obtenir de faibles taux de fausse découverte (les FDRs), dus à la correction pour les tests multiples (lignes oranges). Des régions peuvent montrer un enrichissement substantiel de reads, mais ne seront pas identifiées comme des pics (ligne verte).

3.15.9.2. Approches actuelles pour l'analyse de la liaison différentielle dans ChIP-seq

Récemment, quelques outils statistiques ont été développés pour découvrir les régions qui présentent des différences significatives entre deux ensembles de données ChIP-Seq. Ils se basent principalement sur l'approche quantitative. Cette dernière a un coût de calcul plus élevé, mais elle est recommandée car elle fournit une évaluation statistique précise de la liaison différentielle entre les conditions (par exemple, les facteurs de variation p-valeurs ou q-valeurs liés à l'enrichissement en reads).

L'approche quantitative travaille soit avec le nombre de reads (par exemple, DBChIP (Liang, 2012)) ou sur les densités de reads calculées sur les régions de pics (par exemple, MAnorm (Shao, 2012)). La première citée est une approche intuitive et largement utilisée. Elle est basée sur le rééchantillonnage des données sur la base du nombre total des séquences de reads. Toutefois, cette méthode est inadéquate et peut introduire des erreurs lorsque le ratio signal sur bruit S / N (une abréviation du terme anglais « signal-to-noise ») varie entre les échantillons. On la retrouve notamment dans la méthode de Xu et al. (Xu, 2008), basée sur un modèle de Markov caché, pour détecter les grands domaines de la chromatine associés à des niveaux distincts de modifications d'histones entre deux types de cellules, mais aussi, dans d'autres programmes de peak calling identifiant les régions de liaison différentielles entre deux ensembles de données ChIP-Seq en utilisant un comme échantillon et l'autre comme contrôle (Hongkai, 2008; Rozowsky, 2009 ; Zhang, 2008). Malheureusement, étant donnée leur méthode de base, ils ne parviennent pas à éviter les problèmes associés à différents ratios de S/N .

Pour contourner ce problème des différences dans le ratio S / N entre les échantillons, d'autres méthodes ont posé des hypothèses (de ce fait, il est fortement conseillé de vérifier que les données répondent aux exigences du logiciel choisi pour l'analyse). Par exemple, DIME (Taslim, 2011) suppose qu'une proportion importante des pics est commune aux conditions sous comparaison. De son côté, MAnorm suppose que les

pics qui sont communs dans les deux conditions ne changent pas de façon significative et qu'ils pourraient servir de référence pour construire le modèle de mise à l'échelle "rescaling" pour la normalisation. Cette approche est basée sur l'hypothèse empirique que si une protéine associée à la chromatine a un nombre important de pics partagés entre deux conditions, la liaison dans ces régions communes auront tendance à être déterminées par des mécanismes similaires, et devraient donc présenter des intensités de liaison globales similaires entre les échantillons. D'autres méthodes par contre reposent sur l'hypothèse qu'un nombre constant de pics est prévu dans les conditions (Bardet, 2011). Dans ce cadre, Taslim et al (Taslim, 2009) ont proposé une méthode non linéaire qui utilise la régression localement pondérée (LOWESS) pour la normalisation des données ChIP-Seq. L'hypothèse sous-jacente de cette méthode est que la distribution à l'échelle du génome des densités de reads a une variance et une valeur moyenne égale entre les échantillons (Taslim, 2009). Le problème potentiel avec cette approche est que la symétrie globale sera introduite après la normalisation, une hypothèse qui ne peut être valable lorsque l'on compare des échantillons biologiques avec des nombres différents de sites de liaison. En outre, cette méthode normalise les échantillons en fonction de la différence absolue des nombres de reads au lieu de ratio \log_2 , couramment utilisé dans les méthodes traditionnelles de MA plot (Smyth, 2005). Donc les différences déduites par cette méthode ne peuvent être utilisées directement pour la comparaison quantitative avec d'autres observations de signification biologique, tels que fold changes dans l'expression de gènes.

D'autres exemples d'outils sont présentés dans le Tableau 10. Certains sont plus performants que d'autres en fonction de la protéine ChIPée (par exemple, ChIPDiff (Xu, 2008) spécialisé pour les marques des histones et POLYPHEMUS (Mendoza-Parra, 2012) pour l'ARN Pol II).

Table 10 : Logiciels pour l'analyse de la liaison différentielle dans ChIP-seq (Bailey, 2013).

Logiciel	Disponibilité	Notes
ChIPDiff	http://cmb.gis.a-star.edu.sg/ChIPSeq/paperChIPDiff.htm	Sites différentiel de modification d'histone en utilisant le modèle de Markov caché
Comparative ChIP-seq	http://www.starklab.org/data/bardet_natprotoc_2011/	ratio Fold change entre les hauteurs de pics normalisées
DBChIP	http://pages.cs.wisc.edu/~kliang/DBChIP/	Assigne des mesures incertaines dans un test de liaison non-différentielle (avec edgeR)
DESeq	http://www.bioconductor.org/packages/release/bioc/html/DESeq.html	Test basé sur un modèle utilisant la distribution binomiale négative
DiffBind	http://www.bioconductor.org/packages/release/bioc/html/DiffBind.html	Analyse de l'affinité de liaison différentielle (utilise edgeR et DESeq)
DIME	http://cran.r-project.org/web/packages/DIME/	Identification différentielle utilisant des mélanges d'ensemble
edgeR	http://www.bioconductor.org/packages/release/bioc/html/edgeR.html	Éstimation empirique de Bayes et tests exacts basés sur la distribution binomiale négative
MACS (version 2)	https://github.com/taoliu/MACS/	Détection de pics différentiels base sur quatre fichiers bedGraph paires
MAnorm	http://bcb.dfc.harvard.edu/~gcyuan/MAnorm/MAnorm.htm	Régression robuste pour dériver un modèle linéaire
MMDiff	http://bioconductor.org/packages/release/bioc/html/MMDiff.html	Différences dans la forme en utilisant les méthodes Kernel
NarrowPeaks	http://bioconductor.org/packages/release/bioc/html/NarrowPeaks.html	Analyse basée sur la forme de la variation en utilisant PCA fonctionnel
POLYPHEMUS	http://cran.r-project.org/web/packages/polyphemus/	Normalisation non-linéaire sur le profil RNA Pol II

CHAPITRE IV

NOUVELLE APPROCHE BIO-INFORMATIQUE POUR L'ANALYSE DES LIAISONS DIFFÉRENTIELLES ASSOCIÉES À UN FACTEUR DE TRANSCRIPTION.

Dans ce chapitre, nous présentons une nouvelle approche bio-informatique pour l'analyse des liaisons différentielles d'un facteur de transcription entre plusieurs types ou conditions cellulaires. La méthode a été appliquée à des données réelles pour déterminer des sites liés de façon différentielle entre trois lignées cellulaires du cancer du sein.

4.1. Pipeline de la nouvelle approche

L'approche est basée sur une évaluation globale du taux des sites de liaison partagés et des mesures pour les changements quantitatifs, intégrés avec des étapes du pipeline mugqic de génome québec. De façon générale, la méthode peut être résumée en cinq étapes principales :

- Étape de pré-traitement de données
- Étape d'analyse primaire
- Étape d'analyse secondaire
- Étape d'analyse quantitative
- Étape d'analyse tertiaire

4.1.1. Étape de pré-traitement de données

Comme les données de départ peuvent être des lectures (reads) de séquences d'ADN déjà alignées (le plus souvent à une ancienne version du génome d'intérêt), sous forme de fichiers Bam (Binary alignment map), la forme binaire des fichiers Sam (Sequence Alignment map), un nettoyage et un pré-traitement des données sont nécessaires (voir partie pré-traitement sur la Figure 22). En effet, une analyse ChIP-seq requiert des lectures brutes associées à des chaînes de caractère traduisant la qualité du séquençage (basecalling), généralement sous un format de fichiers appelé format Fastq. Le but est de collecter les informations concernant la provenance des lectures, tels que la marque de la machine de séquençage, l'encodage de la qualité, etc.

4.1.2. Étapes d'analyse primaire, secondaire et tertiaire

Une fois les fichiers fastq obtenus, nous avons configuré et lancé le pipeline ChIP-seq de Génome Québec. Ce pipeline, appelé mugqic, peut être subdivisé en trois grandes parties: une partie, dite analyse primaire, qui consiste en un ensemble d'opérations dont le but est de réaligner les lectures, obtenues par la partie de pré-traitement, contre la version désirée du génome d'intérêt, de filtrer et de nettoyer les alignements obtenus (voir Figure 22). Cette partie est commune à différentes autres analyses des données NGS, notamment, le DNA-seq et le RNA-seq. Par contre, la deuxième partie, ou l'analyse secondaire, est spécifique à ChIP-seq puisqu'elle consiste à déterminer (calling) des pics des sites de liaison et non des mutations. Enfin, l'analyse tertiaire est l'ensemble des étapes dont l'objectif est de caractériser et d'annoter les pics, d'analyser les motifs et d'effectuer l'annotation fonctionnelle des

sites de liaison. Lors de l'étape de l'annotation fonctionnelle des sites, les groupes de gènes cibles du FT dans les différentes conditions sont caractérisés en utilisant les collections de Gene Ontology (GO), comme la base de données des signatures moléculaires (MSig DB), incluant les voies métaboliques et celles de signalisation. L'analyse tertiaire est appliquée aux résultats issus de la partie analyse quantitative, décrite ci-dessous (voir aussi la Figure 22).

4.1.3. Étape d'analyse quantitative

L'analyse quantitative consiste en l'ensemble des étapes dont le but est d'identifier le nombre de sites de liaison partagés entre les conditions deux à deux et le nombre de sites spécifiques à chacune d'elles. Elle permet aussi de caractériser les sites des liaisons différentielles et invariantes du FT. Les comparaisons sont toujours faites deux à deux et les sites sont catégorisés dans trois ensembles : les sites des liaisons croissantes (increase) ou décroissantes (decrease) dans une condition par rapport à l'autre, et les sites des liaisons invariantes (invariant) entre deux conditions. Dans ce qui suit, nous présentons les deux principales sous-étapes de l'analyse quantitative.

4.1.3.1. Évaluation globale du taux des sites de liaison partagés

Bien que ce soit approprié de définir de façon rigoureuse un ensemble de pics avec une confiance élevée (utiliser des seuils serrés), cette méthode ne discrimine pas un site de liaison sous le seuil des vraies régions non-liées. En plus, cette dernière est sujette au bruit, qui pourrait mener à une sous-estimation du chevauchement dans la liaison entre deux ensembles de données. Ainsi, elle ne donne pas une bonne

estimation du nombre de pics qui sont partagés entre les conditions par rapport aux pics qui sont spécifiques à une condition particulière. D'un autre côté, l'objectif de l'évaluation des liaisons entre différentes conditions est de trouver le nombre de sites de liaison communs et uniques, un scénario dans lequel les mesures de signification ne doivent pas être corrigées pour le test multiple. En effet, la tâche n'est pas d'évaluer la signification des meilleurs pics partagés, mais plutôt d'évaluer équitablement le nombre des deux types de pics.

Afin de contourner ces problèmes, et dans le but d'augmenter la sensibilité pour la détection des régions différemment liées (au détriment de l'augmentation du nombre de faux positifs), des seuils plus détendus peuvent être utilisés pour trouver des pics dans chaque condition, une idée appliquée avec succès dans une étude de comparaison entre des espèces vertébrées (Bardet, 2011).

Ainsi, nous avons calculé le taux de pics partagés en termes de pourcentage entre des paires de conditions de la façon suivante. D'abord, nous avons identifié les pics dans la condition A avec un seuil strict corrigé pour le test multiple (FDR), par la suite nous avons évalué la liaison dans la condition B sur la base d'un enrichissement non aléatoire de ChIP (non corrigé pour le test multiple) aux positions correspondantes à chaque site de liaison dans l'échantillon A. Comme les résultats diffèrent dépendamment de l'échantillon pris comme référence, l'expérience condition B vs condition A est aussi réalisée. Comme résultat, nous aurons l'ensemble des pics partagés entre les conditions qui est beaucoup plus précis que si nous avons considéré les pics avec un seuil corrigé pour les tests multiples dans les deux conditions à la fois. En effet, prendre des pics de haute qualité comme référence dans seulement une condition à la fois est une façon moins sévère de filtrer les sites de liaison de haute confiance, ce qui permet une meilleure sensibilité pour la détection du nombre de sites de liaison partagés.

Bien que les sites de liaison non partagés soient témoins de liaisons différentielles d'un facteur de transcription, se prononcer sur les sites de liaison partagés est bien moins évident. En effet, les sites partagés présentent des hauteurs de pics variées entre les conditions, pouvant être témoins de liaison différentielle ou pas. Ainsi, l'analyse quantitative est la plus appropriée pour l'établissement des liaisons différentielles d'un facteur de transcription.

4.1.3.2. Évaluation des changements quantitatifs de la liaison du facteur de transcription

La liaison d'un facteur de transcription à travers une population de cellules n'est pas un phénomène tout ou rien, mais représente plutôt une mesure quantitative (Toth, 2000), c'est à dire, des facteurs de transcription peuvent occuper leurs sites de liaison à des taux différents. Pour mesurer ces différences de liaison quantitative à travers des échantillons, les changements quantitatifs dans les hauteurs de pics peuvent être analysés (Bradley, 2010).

Ainsi, nous avons d'abord identifié les pics de haute confiance pour chacune des conditions, indépendamment, en utilisant un seuil corrigé pour tests multiples FDR (1%). Ensuite, nous avons effectué des comparaisons deux à deux. Pour une comparaison d'une paire de condition A et B, nous avons compilé un ensemble unique de pics provenant de la condition A et B. Les positions des pics issues de l'union des deux ensembles sont ensuite utilisées pour évaluer les hauteurs de pics dans les positions génomiques correspondantes dans les conditions A et B. Ainsi, les hauteurs de pic sont évaluées, même lorsque le pic n'a pas été appelée dans une condition spécifique. Comme des pics de toutes les conditions sont analysés, les

changements identifiés dans la liaison entre les conditions sont intrinsèquement sans biais (c'est à dire, symétrique) par rapport au choix de l'échantillon de référence.

4.2. Application de la méthode aux données du cancer du sein

Pour démontrer l'efficacité et la précision que permet notre pipeline, nous l'avons appliqué sur des données provenant de trois lignées cellulaires du cancer du sein: ER-MCF7 sensibles aux médicaments de l'hormonothérapie, ER-LCC2 résistantes au tamoxifène et ER-LCC9 résistantes au tamoxifène et au fulvestrant.

4.2.1. Contexte de l'étude

Les médicaments utilisés dans l'hormonothérapie du cancer du sein, comprenant les antioestrogènes (le tamoxifène et le fulvestrant) et les inhibiteurs de l'aromatase, bloquent toute activité de ER alpha (se référer à la section 1.1.5.3 du chapitre I). Malheureusement, malgré le profil favorable et l'efficacité de ces agents, l'utilisation de l'hormonothérapie est limitée par l'apparition de la résistance aux médicaments, dans laquelle la plupart des patients qui répondent initialement au traitement peuvent rechuter.

Bien que la perte de ER alpha favorise la résistance aux anti-estrogènes (AE), une grande partie des modèles cellulaires et des tumeurs chez des patients, répondants ou non à la thérapie hormonale, expriment toujours ER alpha et démontrent un recrutement différentiel de ce dernier au niveau de la chromatine (Metzker, 2010). D'un autre côté, la résistance à l'hormonothérapie est fréquente dans les tumeurs ER alpha (+) qui sur-expriment le récepteur du facteur de croissance ERBB2/HER2. Cette surexpression mène à la phosphorylation et à l'activation de ER alpha par les effecteurs de HER2 agissant en aval, incluant AKT et MAPK (Ansorge, 2009;

Kircher, 2011), et ce même en l'absence de l'estrogène (E2) (voir activation indépendante du ligand dans la section 1.1.5.4 du chapitre I). Fait important, l'activation de cette voie induit le recrutement de ER alpha à des sites de la chromatine distincts de ceux induits sous stimulation estrogénique (E2). De façon intéressante, des études ont montré que les changements dans les liaisons ER alpha à l'échelle du génome sont aussi accompagnés par un programme de transcription différent, souvent menant à l'expression constitutive de la majorité des gènes cibles de ER alpha dans les cellules résistantes aux AE (Schuster, 2008 ; Solomon, 1988; Hurtado, 2008). Par ailleurs, ce programme de transcription est aussi associé avec une signature de gènes du cancer du sein HER2 (+) résistant au tamoxifène distincte de celle induite par E2 (Metzker, 2010). Ceci fournirait une explication moléculaire potentielle pour la résistance à l'hormonothérapie dans les cancers du sein ER alpha (+). En effet, ces résultats suggèrent un rôle central pour ER alpha dans les tumeurs du sein résistantes aux hormones qui dépendrait de l'activation des voies de facteurs de croissance, et favoriseraient ainsi le développement de stratégies thérapeutiques antagonisant complètement ER alpha au lieu de bloquer seulement sa réactivité à l'estrogène (Lupien, 2008).

4.2.2. Objectif de l'étude

Dans cette étude, nous nous intéressons aux mécanismes moléculaires conduisant à la résistance aux médicaments dans les cellules du cancer du sein hormono-dépendant par la compréhension du programme de transcription sous-jacent, régulé par le cistrome ER alpha. L'objectif est d'identifier les groupes de gènes, cibles de ER alpha, dans les cellules sensibles versus les cellules résistantes à l'hormonothérapie. Nous caractériserons ces groupes en utilisant les collections de Gene Ontology (GO),

comme la base de données des signatures moléculaires (MSig DB) incluant les voies métaboliques et celles de signalisation. Cette analyse sera basée sur les données ChIP-seq des sites de liaison de ER alpha dans trois lignées cellulaires : ER- MCF7 sensibles aux médicaments, ER-LCC2 résistantes au tamoxifène, et ER-LCC9 résistantes aux tamoxifène et au fulvestrant. Ainsi, nous devons traiter et analyser des quantités importantes de petits fragments, tâche qui ne peut être assurée que par des moyens bio-informatiques.

4.2.3. Problématique et motivation

La détection des régions enrichies, la recherche des motifs sur-représentés et l'identification des motifs sont en général trois étapes indépendantes dans une analyse ChIP-Seq. Ceci peut poser des problèmes de compatibilité liés aux formats de fichier propres à chacun des programmes choisis. Ainsi, pour passer d'un logiciel à un autre, il faut à chaque fois parcourir les fichiers pour récupérer les sorties et les organiser dans le format accepté par le programme suivant. Tout ceci sans oublier la perte d'information qui peut se produire en passant d'une étape à l'autre. Toutefois, des pipelines regroupant plusieurs de ces étapes existent, mais ces méthodes traditionnelles ne peuvent pas répondre à nos besoins. En effet, elles ne peuvent analyser qu'un seul signal ChIP-seq à la fois et sont incapables d'effectuer un "peak calling" différentiel (c'est le cas du pipeline mugqic de génome québec). Comme nous l'avons décrit dans le chapitre précédent, récemment, quelques approches basées sur la combinaison de peak callers avec des tests statistiques pour détecter l'expression digitale différentielle ont été proposées. Cependant, ces méthodes ne parviennent pas à détecter les changements détaillés dans les liaisons protéine-ADN. De plus, il manquerait toujours à ces méthodes quelques logiciels en aval pour des analyses complémentaires, tels que l'identification des motifs et les annotations des sites de liaisons différentielles retrouvés. Cela pourrait être difficile à effectuer du

fait, encore une fois, du format de sortie, souvent un fichier texte, HTML ou encore PDF. Pour ces raisons, nous présentons ici une méthode bio-informatique originale, simple et efficace pour l'identification, l'analyse et la comparaison des sites de liaison des facteurs de transcription entre des ensembles de données ChIP-Seq.

4.2.4. Méthodologie

Pour répondre à la problématique, il faudrait d'abord identifier et analyser les sites de liaison du facteur de transcription ER alpha, individuellement, dans chaque condition. Ces sites correspondent aux régions enrichies (les pics), suite à l'alignement des lectures (les reads), issues de chip-seq, sur le génome. Ensuite, il faudrait déterminer le taux des sites de liaison partagés entre les conditions deux à deux, ainsi que le taux des sites spécifiques à chaque condition. Ceci permet d'avoir une idée globale des changements dans les régions liées, à savoir, le nombre de sites de liaison conservés, perdus ou acquis d'une condition à une autre. Comme les sites partagés ne veulent pas forcément dire des liaisons identiques, il est essentiel d'analyser la densité des reads au niveau de chaque site de liaison afin de caractériser les changements significatifs, témoins de liaisons différentielles de ER alpha. En plus, il est nécessaire de faire une recherche et une identification des motifs enrichis au niveau de tous ces sites. Enfin, il faudrait associer les régions liées par ER alpha aux gènes cibles afin d'établir la signature de gènes dans chaque condition. En exploitant les ontologies de gènes, il est aussi intéressant de trouver et de comparer les fonctions biologiques, enrichies entre conditions, à la recherche de relation spécifique.

Dans les sections suivantes, une description détaillée des méthodes et matériel utilisés est présentée.

4.2.5. Matériels et Méthodes

La Figure 26 résume les cinq principales étapes d'analyse dont les trois assurées par le biais du pipeline ChIP-seq de Génome Québec. Les deux répliquas de nos trois lignées cellulaires sont analysées en parallèle. Différentes instances du pipeline ont été roulées en changeant à chaque fois le design de l'expérience (Permutation des échantillons) dans le but de faire une analyse comparative.

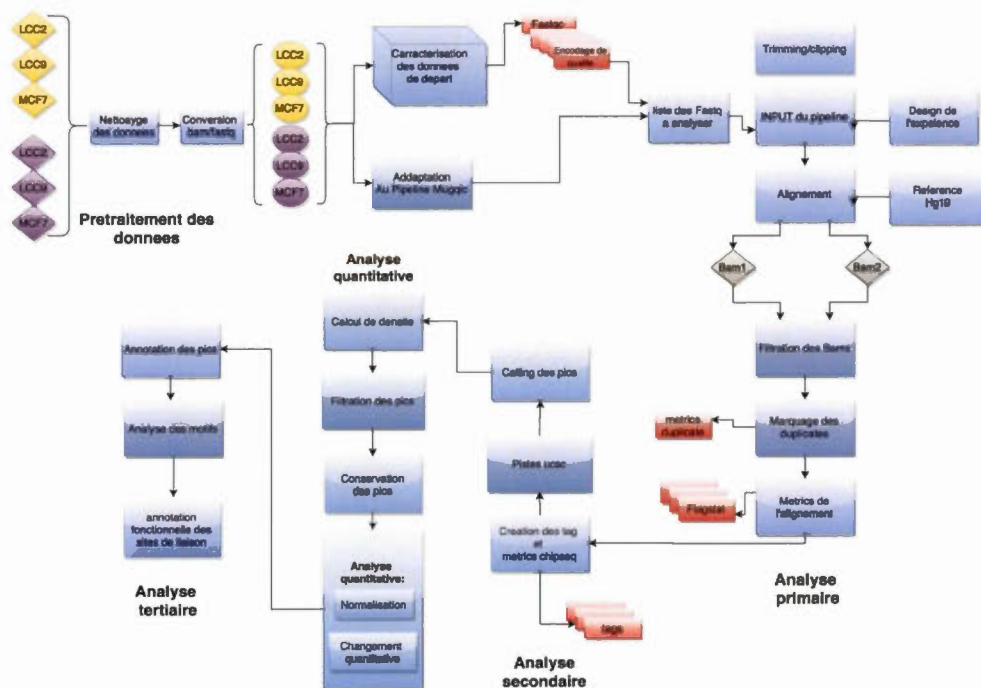


Figure 26 : Pipeline de la nouvelle approche proposée

Nous décrivons dans ce qui suit le déroulement de chacune des étapes du pipeline.

4.2.5.1. Étape de prétraitement des données

Étant donné que nos lectures se trouvent sous forme de séquences préalablement alignées à une ancienne version du génome humain, la première étape de notre analyse a pour but de revenir au format de fichier qui représente le standard *de facto* pour le stockage des sorties des séquenceurs nouvelle génération : le format fastq (Cock, 2010) .

Le format fastq est un format de fichier permettant de stocker, à la fois et sous forme de texte, des séquences d'ADN ou d'ARN et les scores de qualité qui y sont associés. La séquence et le score sont codés chacun avec un seul caractère ASCII. Ce format est un fichier texte qui n'a pas d'extension standard (on peut utiliser .fastq). Le Fast Q est adapté du format FASTA, utilisé pour décrire des séquences. Il est émis par le programme GERALD, contenu dans le pipeline d'analyse d'Illumina. Il contient des informations sur le séquenceur, la séquence des lectures, et des indications sur leur qualité (Voir Figure 27). Les données de départ sont des fichiers Bams : des alignements de lectures sur le génome humain hg18. Les fichiers bams contiennent, sous forme binaire, des informations à propos de l'alignement mais aussi les séquences biologiques et les séquences de qualité correspondantes. Il est donc possible de convertir un fichier bam en un fichier fastq sans perte d'information.

```

@GA8-EAS671_0008_FC:7:1:1186:944#0/1

NGACATTTGCTCTTTGTATGTTACTATAAATTTCCATGATAAACTTGAA
AACCCCATGCATGACAGGGTTGGCGAAAGACATTACAAAGTCCATTGTG
CT

+GA8-EAS671_0008_FC:7:1:1186:944#0/1

BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

```

Figure 27 : Exemple d'un fichier de type fastQ, contenant des données de séquençage.

Ligne 1 : débute par @. Elle donne l'identifiant de la séquence et diverses descriptions comme :

GA8-EAS671_0008_FC	Nom unique de l'instrument
7	Ligne sur la Flowcell
1	Numéro du « tile » ou région au sein de la ligne de la flowcell
1186	'coordonnée-x' du cluster à l'intérieur du tile
944	'coordonnée-y' du cluster à l'intérieur du tile
#0	Numéro d'index si séquençage multiplex (0 sinon)
/1	Numéro de lecture /1 ou /2 si séquençage paired-end

Ligne 2: séquence de la lecture.

Ligne 3: débute normalement par +, suivi de l'identifiant donné à la ligne 1.

Ligne 4: code donnant la qualité de la séquence, elle contient le même nombre de signes que la séquence.

- Nettoyage des données : suppression des alignements secondaires

Lors de l'alignement, l'emplacement correct d'une lecture peut être parfois ambigu, par exemple en raison de répétitions. Dans ce cas, il peut y avoir plusieurs alignements pour la même séquence. L'un d'eux est dit primaire tandis que tous les autres sont secondaires. Typiquement, l'alignement désigné primaire est le meilleur alignement, mais la décision peut être arbitraire. De ce fait, en convertissant nos bams au format fastq, on risque d'avoir plusieurs versions du même read sans que ceux-ci aient été générés initialement par le séquenceur. Dans ce cas, le taux de doublons (reads répétitifs), qui est normalement autour de 20 % pour les séquenceurs Illumina, risque d'augmenter de façon significative. Toutefois, on peut profiter du fait que le format sam/bam distingue entre les alignements primaires et secondaires par des "flag" différents. Ainsi, on a filtré nos six bam en appelant une instance du logiciel samtools (Li, 2009). Samtools étant installé sur le serveur guilimin, il a suffi de charger le lien vers l'exécutable en mémoire en utilisant la commande "module load". Le module **view**, permettant de lire les fichiers bam sous forme binaire, a été lancé en spécifiant le paramètre **-F**. Ce dernier sert à ne garder que les reads sans le "flag" **0x0100** (ce flag correspond aux reads qui sont considérés secondaires "non primary alignment").

```
[aouza123@lg-1r17-n02 ~]$ module load muggic/samtools/0.1.18
[aouza123@lg-1r17-n02 ~]$ for i in {LCC2-1,LCC2-2,MCF7-1,MCF7-2,LCC9-1,LCC9-2}; do
samtools view -f 0x100 ${i}.bam > ${i}_filtered.bam ; done
```


Une boucle "for" en bash (voir ci-dessus) prend en entrée les 6 fichiers bam bruts, exécute la commande samtools et génère 6 fichiers bam filtrés ne contenant aucun alignement secondaire.

- Conversion des bams en fastq

Afin de convertir les bams filtrés, une boucle "for" appelle le module SamToFastq.jar de l'outil picard (<http://picard.sourceforge.net>). Ce dernier devait donner deux fichiers fastq en sortie : un pour chaque paire dans le cas d'un séquençage paired-end. Dans notre cas, picard ne parvient à créer que le fichier fastq R1, ce qui laisse entendre qu'on a affaire à du séquençage single-end. Afin de confirmer que c'est bel et bien du single-end, on a calculé les statistiques d'alignement à l'aide de samtools flagstat. Les résultats confirment l'information fournie par picard vue qu'il n'y a aucun read retrouvé "paired" avec un autre. Ainsi, on supprime les fichiers R2 vides et on supprime le "R1" dans la nomenclature du fichier R1.

```
[aouza123@lg-1r17-n04 ~]$ module load muggic/picard/1.100
[aouza123@lg-1r17-n04 ~]$ for i in {LCC2-1,LCC2-2,MCF7-1,MCF7-2,LCC9-1,LCC9-2}; do java
-jar /software/areas/genomics/phase2/software/picard/picard-tools-1.100/SamToFastq.jar
INPUT=${i}_filteres.bam FASTQ=${i}_R1.fastq SECOND+END_FASTQ= ${i}_R2.fastq > ${i}_filtered
.bam ; done
```

Avant de procéder à l'alignement contre la nouvelle version du génome humain hg19, on a investigué en profondeur les fichiers fastq résultants pour mieux déterminer leur origine et avoir une idée sur leur qualité.

- Caractérisation des données de départ

L'analyse en aval des données demande une connaissance préalable de la nature des séquences ainsi que la technologie de séquençage avec laquelle ils ont été générés.

Il est aussi important de connaître la taille des séquences et leur encodage de qualité.

- Encodage de qualité

Dans un fichier Fastq, à chaque base codée par un caractère unique (A, C, G ou T) doit correspondre un score devant être, lui aussi, codé par un caractère unique. L'utilisation du code ASCII permet de surmonter cette contrainte en faisant correspondre à un caractère son code en base dix. Ce code correspond alors au score de qualité (Cock, 2010).

Selon la technologie de séquençage, l'encodage de qualité peut être du phred 33 ou du phred 64. Un score de qualité Q est un entier relié de façon logarithmique à une probabilité d'erreur p . Cette probabilité d'erreur est calculée lors de l'identification d'une base et correspond à la probabilité qu'il s'agisse d'une mauvaise identification. Selon la technologie de séquençage, plusieurs différentes équations ont été utilisées pour calculer ce score de qualité. Bien que les valeurs de score soient identiques au niveau de l'asymptote verticale correspondant aux scores les plus élevés, elles diffèrent concernant les scores les plus faibles (c'est-à-dire pour des probabilités $p > 0.05$ correspondant à des scores $Q < 13$), d'où l'intérêt de connaître l'origine des données. En effet, nos données étant relativement anciennes (alignés contre la version hg18 du génome qui date d'avant 2011), on a besoin d'en connaître l'encodage pour bien pouvoir interpréter leur qualité.

Pour ce faire et dans le but de vérifier le format de qualité des fastq, on a écrit le script **guess_phred.py** (voir ci-dessous) qui s'inspire du code python téléchargé à partir de <https://github.com/brentp/bio-playground/blob/master/reads-utils/>.

```

"""
    awk 'NR % 4 == 0' your.fastq | python %prog [options]

guess the encoding of a stream of qual lines.
"""
import sys
import optparse

RANGES = {
    'Sanger': (33, 93),
    'Solexa': (59, 104),
    'Illumina-1.3': (64, 104),
    'Illumina-1.5': (67, 104)
}

def get_qual_range(qual_str):
    """
    >>> get_qual_range("DLXYXXRXWYYTPMLUUQWTXTRSXSWMDMTRNDNSMJFFFRMV")
    (68, 89)
    """

    vals = [ord(c) for c in qual_str]
    return min(vals), max(vals)

def get_encodings_in_range(rmin, rmax, ranges=RANGES):
    valid_encodings = []
    for encoding, (emin, emax) in ranges.items():
        if rmin >= emin and rmax <= emax:
            valid_encodings.append(encoding)
    return valid_encodings

```

```

def main():
    p = optparse.OptionParser(__doc__)
    p.add_option("-n", dest="n", help="number of qual lines to test default:-1"
                " means test until end of file or until it is possible to "
                " determine a single file-type",
                type='int', default=-1)

    opts, args = p.parse_args()
    print >>sys.stderr, "# reading qualities from stdin"
    gmin, gmax = 99, 0
    valid = []
    for i, line in enumerate(sys.stdin):
        lmin, lmax = get_qual_range(line.rstrip())
        if lmin < gmin or lmax > gmax:
            gmin, gmax = min(lmin, gmin), max(lmax, gmax)
            valid = get_encodings_in_range(gmin, gmax)
            if len(valid) == 0:
                print >>sys.stderr, "no encodings for range: %s" % str((gmin, gmax))
                sys.exit()
            if len(valid) == 1 and opts.n == -1:
                print "\t".join(valid) + "\t" + str((gmin, gmax))
                sys.exit()

        if opts.n > 0 and i > opts.n:
            print "\t".join(valid) + "\t" + str((gmin, gmax))
            sys.exit()

    print "\t".join(valid) + "\t" + str((gmin, gmax))

if __name__ == "__main__":
    import doctest
    if doctest.testmod(optionflags=doctest.ELLIPSIS | \
                        doctest.NORMALIZE_WHITESPACE).failed == 0:
        main()

```

Le script n'ayant pas pu reconnaître le codage de qualité, l'alternative était de faire une analyse manuelle des séquences de qualité. Vu les caractères présents dans l'encodage de qualité, les 100 premières lignes montrent que c'est très probablement du Illumina miseq ou hiseq phred 64 ou bien du Solexa/Illumina phred 64. Compte tenu que l'alignement a été fait versus la version hg18 du génome et que la

technologie miseq est très récente, il s'agirait probablement de données générées par un séquenceur Illumina/Solexa ou illumina hiseq avec un encodage phred 64.

- Taille des lectures

Le script **guess_phred.py** n'a pu fonctionner correctement à cause de la taille des séquences qui est de 40 pb (10 bases ont probablement été supprimées après clipping). On note aussi que la taille des reads est toujours de 40 pb, ce qui est étrange vue que le trimming des adaptateurs ne donne pas une distribution de taille aussi uniforme dans les fichiers fastq de sortie. En effet, on aurait dû avoir des reads de tailles différentes (entre 40 pb et 50 pb, dans le cas où l'adaptateur est de longueur de 10 pb, la longueur la plus usuelle des adaptateurs).

Ainsi, on a pu conclure que nos données sont de type single-end issues d'un séquenceur Illumina/Solexa ou illumina avec un encodage de qualité phred 64.

- Initialisation et lancement du pipeline chipseq

SGE (Sun Grid Engine) est un système de gestion des tâches de calcul sur une machine multiprocesseur ou un réseau de stations de travail. Son rôle essentiel est de permettre à plusieurs utilisateurs simultanés une utilisation efficace des ressources de calcul (processeurs et mémoire) des noeuds d'une machine multiprocesseur à travers une interface usager simplifiée. SGE est un système logiciel gratuit, produit et supporté par Sun Microsystems.

Les tâches distribuées sont soumises au système SGE à l'aide de la commande qsub en utilisant des scripts de soumission. Ces scripts sont utilisés comme des environnements pour la définition de la ligne de commande qui lance effectivement la tâche sur le noeud (ou les noeuds) choisi (s) par qsub.

Le pipeline ChIP-Seq de génome Québec (mugqic) consiste en un programme écrit en python, qui génère des commandes qsub organisées sous forme d'un script bash, destiné à rouler sur un cluster linux de calcul parallèle (le serveur guillimin dans notre cas) à l'aide du système SGE.

La commande qsub permet de soumettre des tâches de calcul au système de batch en séquentiel ou en parallèle, selon des critères configurables par l'utilisateur. Dans notre cas, le système de batch est séquentiel vue l'interdépendance des tâches du pipeline ChIP-seq. En effet, l'alignement, par exemple, ne peut démarrer que si le trimming est terminé.

La commande qsub retourne immédiatement, sans erreur, si la tâche a été acceptée par le système de batch. La tâche ne sera pas obligatoirement exécutée immédiatement, par contre, elle sera certainement exécutée, même si l'utilisateur ferme la connexion à la machine parallèle directement après avoir lancé la commande qsub. Une fois qu'une tâche soumise est en exécution, les sorties "standard output" et "standard error" seront redirigées vers des fichiers localisés dans le répertoire de base de l'utilisateur. Ceux-ci portent des noms uniques constitués ainsi: nom_de_la_tache.oNN, nom_de_la_tache.eNN, où "nom_de_la_tache" est le nom du script de tâche soumis à qsub, "o" signifie standard output, "e" signifie standard error et "NN" est le numéro d'identification de la tâche, alloué par le système de batch. L'initialisation du pipeline ChIP-seq se fait en exécutant le chispeq.py qui prend en argument trois fichiers d'entrée :

Le fichier Readset

Le fichier Readset contient des informations concernant les fastq, telles que l'adresse des fichiers, l'encodage de qualité (voir paragraphe précédent) ainsi que le nom des échantillons correspondant aux fichiers.

Le fichier *design*

Le fichier *design* est un fichier qui contient une ligne par échantillon et deux colonnes. Chacune des colonnes correspond au type du “peak calling”: Narrow/Broad. Les valeurs 1 correspondent aux échantillons “traitement” et 2 aux échantillons “contrôle”.

Selon la nature de l'expérience, on peut avoir deux ou plusieurs échantillons dans chaque fichier.

Sample	Broad, B	Narrow, N
ERC_LCC2-1q10	1	1
ERC_LCC2-1q20	2	2

Le fichier d'initialisation *guillimin.ini*

Le fichier *guillimin.ini* contient les adresses des exécutables des outils appelés pendant l'exécution du pipeline en utilisant la commande `module load`, les différents paramètres de ces outils ainsi que les fichiers de référence indexés tels que le fichier *fasta* de la version hg19 du génome humain.

```

[DEFAULT]
# Cluster
cluster_submit_cmd=qsub
cluster_submit_cmd_suffix= | grep "[0-9]"
cluster_walltime=-l walltime=24:00:0
cluster_cpu=-l nodes=1:ppn=1
cluster_other_arg=-m ae -M $JOB_MAIL -W umask=0002
cluster_queue=-q sw
cluster_work_dir_arg=-d
cluster_output_dir_arg=-j oe -o
cluster_job_name_arg=-N
cluster_cmd_produces_job_id=true
cluster_dependency_arg=-W depend=afterok:
cluster_dependency_sep=:
cluster_max_jobs=30000
tmp_dir=/lb/scratch/

# Modules
module_bwa=mugqic/bwa/0.7.10
module_homer=mugqic/homer/4.7
module_java=mugqic/java/openjdk-jdk1.7.0_60
module_macs2=mugqic/MACS2/2.1.0.20140616
module_mugqic_R_packages=mugqic/mugqic_R_packages/1.0.2
module_mugqic_tools=mugqic/mugqic_tools/2.0.3
module_perl=mugqic/perl/5.18.2
module_picard=mugqic/picard/1.123
module_python=mugqic/python/2.7.8
module_R=mugqic/R-Bioconductor/3.1.2_3.0
module_samtools=mugqic/samtools/0.1.19-gpfs
module_trimmomatic=mugqic/trimmomatic/0.32
module_weblogo=mugqic/weblogo/2.8.2

# Genome
scientific_name=Homo_sapiens
assembly=hg19
assembly_dir=$MUGQIC_INSTALL_HOME/genomes/species/%(scientific_name)s.%(assembly)s
genome_fasta=%(assembly_dir)s/genome/%(scientific_name)s.%(assembly)s.fa
genome_dictionary=%(assembly_dir)s/genome/%(scientific_name)s.%(assembly)s.dict
genome_bwa_index=%(assembly_dir)s/genome/bwa_index/%(scientific_name)s.%(assembly)s.fa

java_other_options=-XX:ParallelGCThreads=1 -Dsamjdk.use_async_io=true -Dsamjdk.buffer_size=4194304

[picard_sam_to_fastq]
ram=10G
cluster_cpu=-l nodes=1:ppn=3

```



```

[trimmomatic]
ram=2G
threads=1
trailing_min_quality=30
min_length=50
adapter_fasta=$MUGQIC_INSTALL_HOME/software/mugqic_pipeline/v1.3/lib/adapters-truseq.fa
illumina_clip_settings=:2:30:15
# To keep overlapping pairs use the following
# illumina_clip_settings=:2:30:15:8:true
cluster_walltime=-l walltime=24:00:0
cluster_cpu=-l nodes=1:ppn=1

[bwa_mem]
other_options=-M -t 11
sequencing_center=McGill University and Genome Quebec Innovation Centre

[picard_sort_sam]
ram=54G
max_records_in_ram=13500000

[bwa_mem_picard_sort_sam]
cluster_cpu=-l nodes=1:ppn=12

[samtools_view_filter]
min_mapq=20

[picard_merge_sam_files]
ram=1700M
max_records_in_ram=250000
cluster_walltime=-l walltime=35:00:0
cluster_cpu=-l nodes=1:ppn=2

[picard_mark_duplicates]
ram=5G
max_records_in_ram=1000000
cluster_cpu=-l nodes=1:ppn=2
cluster_walltime=-l walltime=48:00:0

[homer_annotate_peaks]
proximal_distance=-2000
distal_distance=-10000
distance5d_lower=-10000
distance5d_upper=-100000
gene_desert_size=100000

[homer_find_motifs_genome]
threads=4
cluster_cpu=-l nodes=1:ppn=4

[gq_seq_utils_report]
design_file=/path/to/design.csv
report_dir=report
## Title for report e.g. <Project Name>
report_title=ChIP-Seq
#report_author=<Author Name>
#report_contact=<author@email.com>

```

Une fois les fichiers d'entrée prêts, on lance le pipeline en choisissant les paramètres nécessaires, tels que les étapes du pipeline qui devraient être lancées.

```
[aouza123@lg-1r17-n04 chipseq]$ module load mugqic/python/2.7.6
[aouza123@lg-1r17-n04 chipseq]$ python chipseq.py -h
usage: chipseq.py [-h] [--help] [-c CONFIG [CONFIG ...]] [--s STEPS]
                  [-o OUTPUT_DIR] [-j {pbs,batch}] [-f] [--clean]
                  [-l {debug,info,warning,error,critical}] [-d DESIGN]
                  [-r READSETS] [-v]

Version: 2.1.0

For more documentation, visit our website: https://bitbucket.org/mugqic/mugqic_pipelines/

optional arguments:
  -h, --help            show this help message and exit
  -c CONFIG [CONFIG ...], --config CONFIG [CONFIG ...]
                        config INI-style list of files; config parameters are
                        overwritten based on files order
  -s STEPS, --steps STEPS
                        step range e.g. '1-5', '3,6,7', '2,4-8'
  -o OUTPUT_DIR, --output-dir OUTPUT_DIR
                        output directory (default: current)
  -j {pbs,batch}, --job-scheduler {pbs,batch}
                        job scheduler type (default: pbs)
  -f, --force            force creation of jobs even if up to date (default:
                        false)
  --clean               create 'rm' commands for all job removable files in
                        the given step range, if they exists; if --clean is
                        set, --job-scheduler, --force options and job up-to-
                        date status are ignored (default: false)
  -l {debug,info,warning,error,critical}, --log {debug,info,warning,error,critical}
                        log level (default: info)
  -d DESIGN, --design DESIGN
                        design file
  -r READSETS, --readsets READSETS
                        readset file
  -v, --version          show the version information and exit

Steps:
1- picard_sam_to_fastq
2- trimmomatic
3- merge_trimmomatic_stats
4- bwa_mem_picard_sort_sam
5- samtools_view_filter
6- picard_merge_sam_files
7- picard_mark_duplicates
8- metrics
9- homer_make_tag_directory
10- homer_make_ucsc_file
11- qc_metrics
12- macs2_callpeak
13- homer_annotate_peaks
14- homer_find_motifs_genome
15- annotation_graphs
16- gq_seq_utils_report
[aouza123@lg-1r17-n04 chipseq]$ python chipseq.py -c chipseq.guillimin.ini -d ./design.csv -w 'pwd' -s 3-12 > toRun.sh
```

Le script **chipseq.py** génère un fichier de type qsub qui contient toutes les commandes nécessaires au lancement des différents outils, en prenant comme

paramètres les variables du fichier de configuration et comme entrée les fichiers fastq du fichier *ReadSet*.

4.2.5.2. Étape d'analyse primaire

- Trimming de qualité et clipping des adaptateurs

Après avoir terminé le séquençage, la première étape de l'analyse est d'évaluer la qualité des reads bruts et de supprimer, de "trimmer" ou de corriger les reads qui ne respectent pas les normes ou les standards définis.

Les données brutes générées par les plates-formes de séquençage sont, en effet, compromises par des artefacts de séquençage telles que les erreurs de base calling et les reads de mauvaise qualité. Ces erreurs sont très fréquentes du fait que les plates-formes sont sensibles à un large éventail de défaillance chimique et instrumentale. De plus, des adaptateurs sont rajoutés aux séquences pendant la préparation des librairies (pour permettre l'amplification PCR), et comme ils sont séquencés au même titre que les fragments d'ADN, ils constituent des artefacts (la contamination par l'adaptateur) devant être éliminés ou "clippés".

De nombreux outils d'analyse en aval ne sont pas en mesure de vérifier les reads de faible qualité. Il est donc nécessaire d'effectuer les tâches de filtrage et de trimming à l'avance afin d'éviter de tirer des conclusions biologiques erronées. En général, ces mesures comprennent la visualisation des scores de qualité des bases et les distributions de nucléotides, le clipping des adaptateurs et le trimming des reads basés sur le score de qualité des bases.

Trimmomatic (Bolger, 2014) est l'outil utilisé par le pipeline muggic pour effectuer les tâches de clipping et de trimminig. Nos données étant issues de fichiers bam, le nettoyage des données de départ, incluant la filtration selon la qualité (Quality trimming) ainsi que la suppression des adaptateurs (adapter clipping), a déjà été fait. Il n'est donc pas nécessaire de le refaire. Par conséquent, on a renommé nos fastq de

départ de façon à ce que le pipeline muggic les considère comme étant déjà “trimmé” et “clippé”.

– Alignement

L’alignement prend en entrée un fichier fastq et le génome de référence hg19. Ainsi, il pourra associer chaque lecture à une location qui se traduit par un nom de chromosome et un entier correspondant à une position sur ce dernier. Le génome de référence étant différent du génome de l’échantillon étudié, un match parfait est rare. En effet, la plupart des alignements contiennent souvent un ou plusieurs “mismatch”. La majorité des outils d’alignement donnent en sortie des fichiers de type bam qui contiennent entre autres les reads, leurs positions par rapport au génome de référence et la qualité de l’alignement (mapping quality). Le pipeline muggic utilise la variante MEM de l’outil bwa (Li, 2009) pour aligner les lectures contre la version hg19 du génome humain. L’outil bwa est une implémentation d’une méthode de réorganisation des données appelée la transformée de Burrows-Wheeler (pour plus de détails, se référer à la section 3.10.2 du chapitre III) dont l’une des utilisations en génomique est l’alignement de courtes séquences contre les longs génomes de référence (3 Go pour le hg19) en un temps relativement court.

Notons ici que la variante MEM de l’outil bwa est désignée pour aligner des séquences dont la taille est supérieure à 100 pb et que nos séquences sont de taille 40 pb. Nous n’avons pas changé d’outil d’alignement, mais ceci pourrait être envisagé dans une tentative d’optimisation de l’alignement pour des analyses ultérieures.

– Filtration des Bams

Une fois générés, les fichiers bam doivent subir une série de modifications destinées à les rendre utilisables par les outils en aval. Les “post mapping modifications” se font dans notre cas à l’aide de deux outils : picard et samtools.

- Tri des fichiers Bam

Afin de faciliter l'accès aux outils utilisant les bams, les reads sont triés par ordre numérique selon leurs positions par l’outil picard/SortSam.jar.

```
#
# JOB: bwa_mem_picard_sort_sam_2_JOB_ID: bwa_mem_picard_sort_sam.readset2
#
JOB_NAME=bwa_mem_picard_sort_sam.readset2
JOB_DEPENDENCIES=$trimomatic_2_JOB_ID
JOB_DONE=job_output/bwa_mem_picard_sort_sam/bwa_mem_picard_sort_sam.readset2.5b2171b1043850cf3cd7af5b7354b32a.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_${TIMESTAMP}.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'bwa_mem_picard_sort_sam.readset2.5b2171b1043850cf3cd7af5b7354b32a.mugqic.done'
module load mugqic/bwa/0.7.10 mugqic/java/openjdk-jdk1.7.0_60 mugqic/picard/1.123 && \
mkdir -p alignment/ERC_LCC2-1q20/readset2 && \
bwa mem \
  -M -t 11 \
  -R '@RG\tID:readset2\tSM:ERC_LCC2-1q20\tCN:McGill University and Genome Quebec Innovation Centre\tPL:Illumina' \
  /software/areas/genomics/phase2/genomes/species/Homo_sapiens.hg19/genome/bwa_index/Homo_sapiens.hg19.fa \
  trim/ERC_LCC2-1q20/readset2.trim.single.fastq.gz | \
java -Djava.io.tmpdir=/lb/scratch/ -XX:ParallelGCThreads=1 -Dsamjdk.use_async_io=true -Dsamjdk.buffer_size=4194304 -Xmx54G -jar $PICARD_HOME/SortSam.jar \
  VALIDATION_STRINGENCY=SILENT CREATE_INDEX=true \
  TMP_DIR=/lb/scratch/ \
  INPUT=/dev/stdin \
  OUTPUT=alignment/ERC_LCC2-1q20/readset2/readset2.sorted.bam \
  SORT_ORDER=coordinate \
  MAX_RECORDS_IN_RAM=13500000
bwa_mem_picard_sort_sam.readset2.5b2171b1043850cf3cd7af5b7354b32a.mugqic.done
)
bwa_mem_picard_sort_sam_2_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND
MUGQIC_STATE=$PIPESTATUS
echo MUGQICexitStatus:$MUGQIC_STATE
if [ $MUGQIC_STATE -eq 0 ] ; then touch $JOB_DONE ; fi
exit $MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:00 -q sw -l nodes=1:ppn=12 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "$bwa_mem_picard_sort_sam_2_JOB_ID $JOB_NAME $JOB_DEPENDENCIES $JOB_OUTPUT_RELATIVE_PATH"
" >> $JOB_LIST
_
```


- Indexage des fichiers Bam

Certains outils requièrent des index .bai pour pouvoir lire les fichiers bam. Le module index de samtools est utilisé par le pipeline mugqic afin de générer l'index.

- Filtration des reads alignés

Les reads alignés avec une valeur de qualité inférieure au seuil de qualité, fourni par l'utilisateur (mapq=0 dans notre cas), sont enlevés des fichiers bams à l'aide du module view de samtools.

```
#-----
# JOB: samtools_view_filter_1_JOB_ID: samtools_view_filter.readset1
#-----
JOB_NAME=samtools_view_filter.readset1
JOB_DEPENDENCIES=$bwa_mem_picard_sort_sam_1_JOB_ID
JOB_DONE=job_output/samtools_view_filter/samtools_view_filter.readset1.66f7fd360a507c3b42a9eee490aa4e50.m
mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_$TIMESTAMP.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'samtools_view_filter.readset1.66f7fd360a507c3b42a9eee490aa4e50.mugqic.done'
module load mugqic/samtools/0.1.19-gpfs && \
samtools view -b -F4 -q 20 \
    alignment/ERC_LCC2-1q10/readset1/readset1.sorted.bam \
    > alignment/ERC_LCC2-1q10/readset1/readset1.sorted.filtered.bam
samtools_view_filter.readset1.66f7fd360a507c3b42a9eee490aa4e50.mugqic.done
)
samtools_view_filter_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND
MUGQIC_STATE=$PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ]; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:
0 -q sw -l nodes=1:ppn=1 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "$samtools_view_filter_1_JOB_ID      $JOB_NAME      $JOB_DEPENDENCIES      $JOB_OUTPUT_RELATIVE_PATH
" >> $JOB_LIST
```

Dans le cas où plusieurs bams correspondent à un seul échantillon, l'outil picard/MergeSam est utilisé par mugqic afin de fusionner les bams correspondants (pas dans notre cas).

```

#-----
# STEP: picard_merge_sam_files
#-----
STEP=picard_merge_sam_files
mkdir -p $JOB_OUTPUT_DIR/$STEP

#-----
# JOB: picard_merge_sam_files_1_JOB_ID: symlink_readset_sample_bam.ERC_LCC2-1q10
#-----
JOB_NAME=symlink_readset_sample_bam.ERC_LCC2-1q10
JOB_DEPENDENCIES=$samtools_view_filter_1_JOB_ID
JOB_DONE=job_output/picard_merge_sam_files/symlink_readset_sample_bam.ERC_LCC2-1q10.b71ed2646a9cb88ceef74d519c85501.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_$TIMESTAMP.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'symlink_readset_sample_bam.ERC_LCC2-1q10.b71ed2646a9cb88ceef74d519c85501.mugqic.done'
mkdir -p alignment/ERC_LCC2-1q10 && \
ln -s -f readset1/readset1.sorted.filtered.bam alignment/ERC_LCC2-1q10/ERC_LCC2-1q10.merged.bam
symlink_readset_sample_bam.ERC_LCC2-1q10.b71ed2646a9cb88ceef74d519c85501.mugqic.done
)
picard_merge_sam_files_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND
MUGQIC_STATE=$PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ] ; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:
0 -q sw -l nodes=1:ppn=1 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "$picard_merge_sam_files_1_JOB_ID $JOB_NAME $JOB_DEPENDENCIES $JOB_OUTPUT_RELATIVE_PATH
" >> $JOB_LIST

```

— Marquage des doublons

En raison d'erreurs inhérentes à la technologie de séquençage, certaines lectures auront des copies exactes issues du même fragment d'ADN. Ces lectures sont appelées doublons PCR et se produisent à cause d'erreurs d'amplification lors de la PCR. Les doublons partagent la même séquence et la même position d'alignement et pourraient causer des problèmes au cours du peak calling en sur-représentant certaines régions. Des doublons peuvent aussi se créer lors du "base calling". En effet, le laser du séquenceur peut capter deux fois le même fragment et créer ce qu'on appelle un doublon optique. Il est important de noter que la probabilité d'avoir des doublons est beaucoup plus grande dans les approches Single-end. Par conséquent,

Picard/MarkDuplicate est utilisé pour marquer les doublons dans les fichiers BAM, en les définissant et en leur attribuant un “flag” spécifique. Les outils en aval reconnaissent ce “flag” et n'utiliseront pas les reads ainsi marqués.

```
#-----
# STEP: picard_mark_duplicates
#-----
STEP=picard_mark_duplicates
mkdir -p $JOB_OUTPUT_DIR/$STEP

#-----
# JOB: picard_mark_duplicates_1_JOB_ID: picard_mark_duplicates.ERC_LCC2-1q10
#-----
JOB_NAME=picard_mark_duplicates.ERC_LCC2-1q10
JOB_DEPENDENCIES=$picard_merge_sam_files_1_JOB_ID
JOB_DONE=job_output/picard_mark_duplicates/picard_mark_duplicates.ERC_LCC2-1q10.07f1370360d1aa18c0b6794aa27e44ce.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_$TIMESTAMP.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'picard_mark_duplicates.ERC_LCC2-1q10.07f1370360d1aa18c0b6794aa27e44ce.mugqic.done'
module load mugqic/java/openjdk-jdk1.7.0_60 mugqic/picard/1.123 && \
java -Djava.io.tmpdir=/lb/scratch/ -XX:ParallelGCThreads=1 -Dsamjdk.use_async_io=true -Dsamjdk.buffer_size=4194304 -Xmx5G -jar $PICARD_HOME/MarkDuplicates.jar \
  REMOVE_DUPLICATES=false CREATE_MDS_FILE=true VALIDATION_STRINGENCY=SILENT CREATE_INDEX=true \
  TMP_DIR=/lb/scratch/ \
  INPUT=alignment/ERC_LCC2-1q10/ERC_LCC2-1q10.merged.bam \
  OUTPUT=alignment/ERC_LCC2-1q10/ERC_LCC2-1q10.sorted.dup.bam \
  METRICS_FILE=alignment/ERC_LCC2-1q10/ERC_LCC2-1q10.sorted.dup.metrics \
  MAX_RECORDS_IN_RAM=1000000
picard_mark_duplicates.ERC_LCC2-1q10.07f1370360d1aa18c0b6794aa27e44ce.mugqic.done
)
picard_mark_duplicates_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND"
MUGQIC_STATE=\PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ] ; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=48:00:
0 -q sw -l nodes=1:ppn=2 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "picard_mark_duplicates_1_JOB_ID $JOB_NAME $JOB_DEPENDENCIES $JOB_OUTPUT_RELATIVE_PATH"
" >> $JOB_LIST
```

– Qc des séquences et métriques de l'alignement

Le cheminement des séquences à travers les étapes précédentes du pipeline génère un ensemble de métriques qui servent à apprécier la qualité des séquences et la qualité de l'alignement :

- Le nombre total des reads

Le nombre total de reads générés par le séquenceur : plus la profondeur du séquençage est grande plus cette valeur est grande. Un séquençage profond permet d'avoir une meilleure précision et sensibilité lors du "calling" des pics.

- Le nombre des reads après trimming

Le rapport entre le nombre des reads "trimmés" et le nombre total des reads est inversement proportionnel à la qualité des séquences. Toutefois, cette valeur peut varier selon l'outil du trimming choisi. Il existe en effet des outils avec un paramétrage agressif par défaut (Fastx) et des outils avec un paramétrage moins agressif (Trimmomatic dans notre cas).

- Pourcentage des reads alignés

Cette valeur reflète directement la qualité de l'alignement, qui, lui même dépend de la qualité des séquences. Ce pourcentage est primordial puisque le nombre des reads alignés au génome influence la qualité de l'analyse ChIP-seq en aval. Cette valeur est produite par l'outil d'alignement bwa ainsi que par le module flagstat de samtools.

- Nombre de reads filtrés selon la qualité de l'alignement

Après l'alignement, les reads sont filtrés selon leur **mapq (mapping quality)** et il est important de connaître le nombre de reads bien alignés avec lesquelles on va travailler par la suite.

- Pourcentage de doublons

La fraction des reads dupliqués n'est pas à considérer parmi les reads ayant réussi le passage de la filtration par mapq puisque les outils en aval ignorent les reads marqués lors de l'étape de marquage de doublons.

```
#-----
# STEP: metrics
#-----
STEP=metrics
mkdir -p $JOB_OUTPUT_DIR/$STEP

#-----
# JOB: metrics_1_JOB_ID: metrics.readset1
#-----
JOB_NAME=metrics.readset1
JOB_DEPENDENCIES=$merge_trimmomatic_stats_1_JOB_ID:$bwa_mem_picard_sort_sam_1_JOB_ID
JOB_DONE=job_output/metrics/metrics.readset1.2fcabb11d1fe831703b13b029c1bf1da.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_$TIMESTAMP.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'metrics.readset1.2fcabb11d1fe831703b13b029c1bf1da.mugqic.done'
module load mugqic/samtools/0.1.19-gpfs mugqic/perl/5.18.2 mugqic/mugqic_tools/2.0.3 && \
samtools flagstat \
  alignment/ERC_LCC2-1q10/readset1/readset1.sorted.bam \
  > alignment/ERC_LCC2-1q10/readset1/readset1.sorted.bam.flagstat && \
perl -MReadMetrics -e 'ReadMetrics::mergeStats(
  "ERC_LCC2-1q10 readset1",
  "metrics/readset1.readstats.csv.tmp",
  ReadMetrics::parseFlagstats(
    "ERC_LCC2-1q10 readset1",
    "alignment/ERC_LCC2-1q10/readset1/readset1.sorted.bam.flagstat"
  )
)' && \
grep "ERC_LCC2-1q10 readset1" metrics/trimming.stats | cut -f3- | sed 's/ /, /g' | sed '1iRaw Fragme
nts,Fragment Surviving,Single Surviving' | paste -d, <(cut -f1 -d, metrics/readset1.readstats.csv.tmp) -
<(cut -f2- -d, metrics/readset1.readstats.csv.tmp) \
> metrics/readset1.readstats.csv && \
rm metrics/readset1.readstats.csv.tmp

metrics.readset1.2fcabb11d1fe831703b13b029c1bf1da.mugqic.done
)
metrics_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND
MUGQIC_STATE=\$PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ] ; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:
0 -q sw -l nodes=1:ppn=1 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "$metrics_1_JOB_ID $JOB_NAME $JOB_DEPENDENCIES $JOB_OUTPUT_RELATIVE_PATH" >> $JOB_LIST
```

4.2.5.3. Étape d'analyse secondaire

– Création des “tag” et métriques Chip-seq

```

#-----
# STEP: homer_make_tag_directory
#-----
STEP=homer_make_tag_directory
mkdir -p $JOB_OUTPUT_DIR/$STEP

#-----
# JOB: homer_make_tag_directory_1_JOB_ID: homer_make_tag_directory.ERC_LCC2-1q10
#-----
JOB_NAME=homer_make_tag_directory.ERC_LCC2-1q10
JOB_DEPENDENCIES=spicard_mark_duplicates_1_JOB_ID
JOB_DONE=job_output/homer_make_tag_directory/homer_make_tag_directory.ERC_LCC2-1q10.aaf67250fa18e8d9be1a1369d27fe1e8.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_$TIMESTAMP.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'homer_make_tag_directory.ERC_LCC2-1q10.aaf67250fa18e8d9be1a1369d27fe1e8.mugqic.done'
module load mugqic/samtools/0.1.19-gpfs mugqic/homer/4.7 && \
makeTagDirectory \
  tags/ERC_LCC2-1q10 \
  alignment/ERC_LCC2-1q10/ERC_LCC2-1q10.sorted.dup.bam \
  -checkGC -genome hg19
homer_make_tag_directory.ERC_LCC2-1q10.aaf67250fa18e8d9be1a1369d27fe1e8.mugqic.done
)
homer_make_tag_directory_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND"
MUGQIC_STATE=$PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ] ; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:
0 -q sw -l nodes=1:ppn=1 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "$homer_make_tag_directory_1_JOB_ID          $JOB_NAME          $JOB_DEPENDENCIES          $JOB_OUTPUT_RELAT
IVE_PATH" >> $JOB_LIST

```

Afin de faciliter l'analyse ChIP-Seq, le logiciel HOMER4 (Heinz, 2010) transforme l'alignement de séquence en une structure de données indépendante de la plate-forme. Cette structure de données est une représentation de l'expérience. Toutes les informations pertinentes concernant l'expérience sont organisées en un répertoire de “Tag” ou balises. Lors de la création de répertoires de balises, plusieurs fonctions de contrôle de qualité sont exécutées pour aider à fournir des informations et des commentaires sur la qualité de l'expérience. Plusieurs paramètres importants sont des estimations qui sont par la suite utilisées pour l'analyse en aval, tels que la longueur approximative des fragments ChIP-Seq.

Quatre mesures de qualité sont obtenues avec le logiciel HOMER:

- Le nombre de “tag” par position

Le nombre de “tag” par position représente le nombre de lectures par position unique. Si une expérience contient un nombre de reads dupliqués trop important, le nombre moyen de “tag” augmente proportionnellement. Un nombre de tag moyen inférieur à 1,5 est généralement recommandé.

- Auto-correlation des “tag”

L’auto-corrélation des tags est la distribution de la distance entre les lectures adjacentes. Deux distributions sont présentées, une pour le brin principale et une pour le brin opposé. La valeur maximale de la distribution du brin opposé représente habituellement la longueur du fragment.

- Le biais de séquence

Le biais de séquence montre la relation entre le read séquencé et la séquence génomique réelle. De petites variations dans le début du read peuvent se produire, mais dans l’ensemble, les ratios de nucléotides doivent rester constants.

- Le biais GC

Le biais GC montre le contenu de GC pour chaque lecture. Si la distribution de GC du read est décalée par rapport à la distribution habituelle du génome, le read pourrait être surreprésenté dans les régions riches ou pauvres en GC.

– Pistes ucsc

```
#-----
# STEP: homer_make_ucsc_file
#-----
STEP=homer_make_ucsc_file
mkdir -p $JOB_OUTPUT_DIR/$STEP

#-----
# JOB: homer_make_ucsc_file_1_JOB_ID: homer_make_ucsc_file.ERC_LCC2-1q10
#-----
JOB_NAME=homer_make_ucsc_file.ERC_LCC2-1q10
JOB_DEPENDENCIES=$homer_make_tag_directory_1_JOB_ID
JOB_DONE=job_output/homer_make_ucsc_file/homer_make_ucsc_file.ERC_LCC2-1q10.072d7faf97a767e71a78b0db8f1a7
aae.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_$TIMESTAMP.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'homer_make_ucsc_file.ERC_LCC2-1q10.072d7faf97a767e71a78b0db8f1a7aae.mugqic.done'
module load mugqic/homer/4.7 && \
mkdir -p tracks/ERC_LCC2-1q10 && \
makeUCSCfile \
    tags/ERC_LCC2-1q10 | \
gzip -1 -c > tracks/ERC_LCC2-1q10/ERC_LCC2-1q10.ucsc.bedGraph.gz
homer_make_ucsc_file.ERC_LCC2-1q10.072d7faf97a767e71a78b0db8f1a7aae.mugqic.done
)
homer_make_ucsc_file_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND
MUGQIC_STATE=\$PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ] ; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:
0 -q sw -l nodes=1:ppn=1 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "$homer_make_ucsc_file_1_JOB_ID      $JOB_NAME      $JOB_DEPENDENCIES      $JOB_OUTPUT_RELATIVE_PATH
" >> $JOB_LIST
```

Des pistes au format bedgraph sont générées par l’outil HOMER à partir des alignements de séquences. Ces pistes peuvent être visualisées sur le “genome browser” de ucsc (Figure 28) et donner ainsi une idée concrète sur la distribution des reads sur le génome de référence, le pourcentage de couverture des régions cibles ainsi que la profondeur du séquençage.

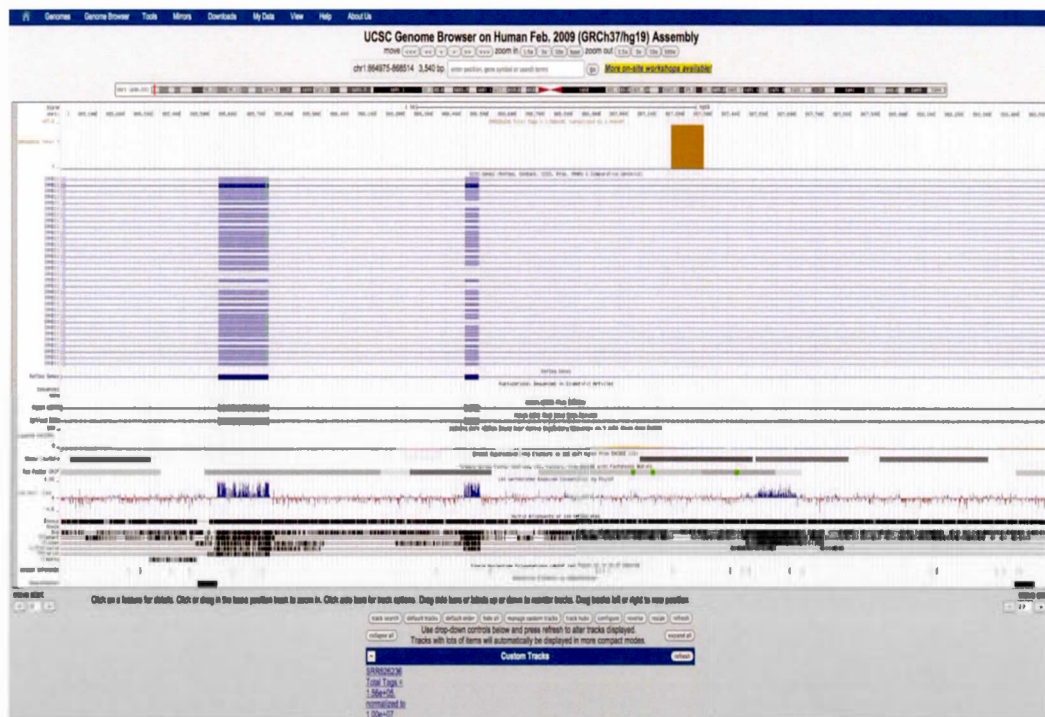


Figure 28 : Piste ucsc de l'expérience ER-LCC2-1_VS_ER-LCC9-2

- Calling des pics

Comme vu précédemment, le "calling" des pics est une méthode de calcul utilisée pour identifier les domaines génomiques enrichis en reads alignés lors de l'exécution d'une expérience ChIP-seq. Ces zones sont celles où une protéine interagit avec l'ADN. Si la protéine est un facteur de transcription, la zone enrichie est le site de fixation de ce facteur de transcription (TFBS).

Le pipeline muggic utilise l'outil MACS (Zhang, 2008), dans sa version 2.1.0, pour faire le "peak calling" (la q-value cutoff = $5.00e-02$). Ce dernier modélise empiriquement la longueur Lambda des fragments ChIP-seq et l'utilise pour améliorer

la résolution spatiale des sites de liaison prédits (le rang pour calculer le lambda régional est : 10000 pbs). Il se base sur une distribution Poisson pour capturer, efficacement et de façon dynamique, les biais locaux dans la séquence du génome, ce qui permet une prédiction plus sensible et plus robuste.

Ne disposant pas d'échantillons contrôles, MACS est tout de même capable de modéliser une distribution servant de background et appeler nos pics. Comme résultats, il produit des fichiers de type bed représentant les pics par leur localisation chromosomique et des métriques inhérentes au "calling". Le fichier NAME_summits.bed, qui contient les sommets et les emplacements pour tous les pics et dont la 5ème colonne représente l'auteur du "pileup" (reads empilés dans une région génomique), est recommandé dans le cas où on recherche des motifs.

```
#-----
# STEP: macs2_callpeak
#-----
STEP=macs2_callpeak
mkdir -p $JOB_OUTPUT_DIR/$STEP

#-----
# JOB: macs2_callpeak_1_JOB_ID: macs2_callpeak.Narrow
#-----
JOB_NAME=macs2_callpeak.Narrow
JOB_DEPENDENCIES=$picard_mark_duplicates_1_JOB_ID:$picard_mark_duplicates_2_JOB_ID
JOB_DONE=job_output/macs2_callpeak/macs2_callpeak.Narrow.75a93ed3d764bfc3c482417b18027e36.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_$TIMESTAMP.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'macs2_callpeak.Narrow.75a93ed3d764bfc3c482417b18027e36.mugqic.done'
module load mugqic/python/2.7.8 mugqic/MACS2/2.1.0.20140616 && \
mkdir -p peak_call/Narrow && \
macs2 callpeak --format BAM --nomodel \
  --gsize 2509729011.2 \
  --treatment \
  alignment/ERC_LCC2-1q20/ERC_LCC2-1q20.sorted.dup.bam \
  --control \
  alignment/ERC_LCC2-1q10/ERC_LCC2-1q10.sorted.dup.bam \
  --name peak_call/Narrow/Narrow \
  >& peak_call/Narrow/Narrow.diag.macs.out
macs2_callpeak.Narrow.75a93ed3d764bfc3c482417b18027e36.mugqic.done
)
macs2_callpeak_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND"
MUGQIC_STATE=$PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ]; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:
0 -q sw -l nodes=1:ppn=1 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "$macs2_callpeak_1_JOB_ID $JOB_NAME $JOB_DEPENDENCIES $JOB_OUTPUT_RELATIVE_PATH" >> $JOB_LIST
```

4.2.5.4. Étape d'analyse quantitative et catégorisation des pics

Dans cette partie, on effectue une catégorisation et une analyse quantitative à partir des pics bruts générés par le pipeline de génome Québec. Cette étape a été inspirée de la littérature, d'une étude CHIP-seq entre plusieurs espèces. L'analyse différentielle s'est faite entre paires de conditions en l'absence d'échantillons de contrôle.

– Calcul de la densité des lectures

```
# Create density file: extend reads, calculate read density at each position and normalize the library size to 1 million reads
# ER_MCF7-1 255 pb
# ER_MCF7-2 271 pb
# ER_LCC2-1 249 pb
# ER_LCC2-2 243 pb
# ER_LCC9-1 247 pb
# ER_LCC9-2 238 pb
EXTEND=150; for sample in MCF7-1 MCF7-2 LCC2-1 LCC2-2 LCC9-1 LCC9-2 ; do bam="ER_${sample}.q10.sorted.dup.bam "; echo $bam
; librarysize=$(samtools idxstats $bam | awk '{total+= $3}END{print total}');
bamToBed -i ${sample}.bam | awk -vCHROM= hg19.chrom.sizes -vEXTEND=$EXTEND -vOFS = '\t'
BEGIN{while(getline>CHROM){chromSize[$1] = $2}}{chrom = $1;start = $2;end = $3;
strand = $6;if(strand == " + "){end = start + EXTEND;if(end>chromSize[chrom]){end = chromSize[chrom]}};if(strand == " -
"){start = end-EXTEND;if(start<1){start = 1}};print chrom,start,end}' | sort -k1,1 -k2,2n | genomeCoverageBed -i stdin -g
hg19.chrom.sizes -d | awk -vOFS='\t' -vSIZE=$librarySize '{print $1,$2,$2+1,$3*1000000/SIZE}' | gzip > ${sample}.densi
ty.gz
```

Dans cette étape, on va créer des fichiers BigWig (la version binaire compressée des fichiers WIG -voir Figure 29-), dits de visualisation de la densité des reads. Le nombre de reads, à la position près ou dans un intervalle, est d'abord calculé. Cependant, il ne faut pas oublier le paramètre d'extension des reads. En effet, à l'origine, la chromatine a été fragmentée jusqu'à ce que la majorité des fragments aient une certaine taille. Par la suite, l'ADN a été sélectionné suivant cette longueur, et seulement une extrémité de ces mêmes fragments a été séquencée. En étendant la

lecture à la longueur moyenne des fragments génomiques, connue a priori ou déterminée durant le peak calling, on cherche à recréer de manière artificielle le fragment d'origine. Le read est étendu de x bases à son extrémité 3' ou « end », selon le brin sur lequel il a été aligné.

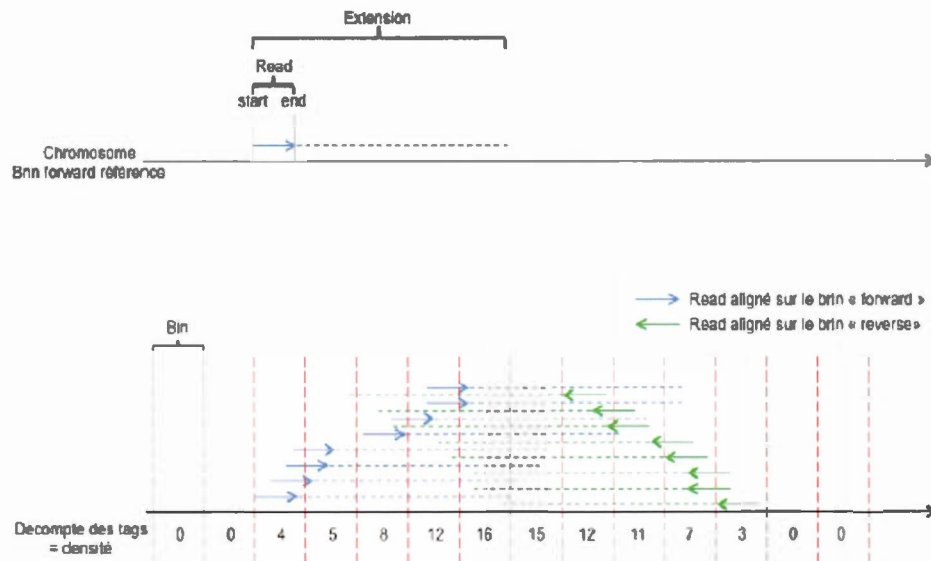


Figure 29: Description de la création d'un fichier de densité WIG (source UCSC)

Tout comme les pistes ucsc, les pistes BigWig peuvent être visualisées dans un "genome browser". Cependant, dans notre cas, le but est surtout de les utiliser ultérieurement lors de l'étape de l'analyse quantitative, ainsi que pour faire une étude de reproductibilité entre les répliquats.

– Filtation des pics et création des pics de contrôle

Les pics déterminés par MACS sont filtrés de façon stricte ($FDR \leq 1\%$) afin d'avoir un ensemble de pics de haute qualité (macs_confident.txt), puis avec la valeur par

défaut ($P\text{-value} < 10^{-5}$) pour identifier les régions avec des enrichissements non aléatoires (macs_enrichment.txt). On crée aussi des pics contrôles en déplaçant des pics à des endroits aléatoires (macs_control.txt). Ces informations sont utiles pour les étapes suivantes.

```
# Print shift d (2*d = genomic fragment length)
for pair in MCF7-1 MCF7-2 LCC2-1 LCC2-2 LCC9-1 LCC9-2; do grep -# d = " ${pair}_macs_p05_peaks.xls | awk '{print $4}'; done
> d.txt

# Number of peaks at different FDR thresholds
echo -e "FDR\tAll\t5\t1\t0" > peaks_number_thr.txt
for pair in MCF7-1 MCF7-2 LCC2-1 LCC2-2 LCC9-1 LCC9-2; for fdr in 100 5 1 0; do
echo -en "\t${grep -v -# "${pair}_macs_p05_peaks.xls | awk -vFDR=$fdr '(NR>166$9>=FDR)' | wc -l) ;done; done
> ${pair}_number_of_peaks_FDR.txt

# Define confident peaks (FDR), enriched regions (P-value>=10e-5) and control peaks FDR=1
FDR=1;
for pair in MCF7-1 MCF7-2 LCC2-1 LCC2-2 LCC9-1 LCC9-2; do
    # Confident peaks
grep -v -# "${pair}_macs_p05_peaks.xls | awk -vOFS='\t' -vFDR=$FDR '(NR>166$9>=FDR){if($2>1){$2=1};print $1,$2,$3,$5,$7,$8,$9}' > ${pair}_macs_confident.txt
    # Regions with significant enrichment
grep -v -# "${pair}_macs_p05_peaks.xls | awk -vOFS='\t' '(NR>1) {if($2>1) {$2=1};print $1,$2,$3,$5,$7,$8,$9}' > ${pair}_macs_enrichment.txt
    # Control peaks
shuffleBed -i ${pair}_macs_enrichment.txt -g hg19.chrom.sizes -chrom | sort -k1,1 -k2,2n > ${pair}_macs_control.txt
done
```

— Conservation des pics

Afin de contourner les problèmes, posés dans d'autres méthodes lors de la détermination des pics conservés (partagés) entre des conditions, on ne va pas faire une simple intersection des régions de pics. L'idée est de calculer un taux de conservation entre les deux conditions *A* et *B*. Ce dernier représente le pourcentage de pics de haute qualité identifiés dans la condition *A* correspondants à des régions enrichies de façon non aléatoire dans la condition *B*. Pour ce faire, on va se servir des résultats obtenus dans l'étape précédente (à savoir les fichiers de liste de pics de haute

qualité et ceux des régions enrichies de façon non aléatoire). Ainsi, bien que les pics dans l'échantillon de référence soient appelés avec un seuil FDR strict (haute qualité), on évalue la liaison dans les autres échantillons par un enrichissement non aléatoire (non corrigé pour le test multiple) aux positions correspondantes, à chaque site de liaison dans l'échantillon de référence.

Afin d'éviter de compter les chevauchements parasites entre les pics, nous avons fait en sorte de respecter une règle : le sommet du pic doit chevaucher la région avec un enrichissement non aléatoire.

Pour ne pas sous-estimer substantiellement la divergence, on calcule aussi la conservation pour l'ensemble des pics de contrôle (c.-à-dire pics transférés à des localisations aléatoires), créé dans l'étape précédente.

L'expérience inverse B vs A est aussi effectuée.

```
#exp1
reference=LCC2-1
sample=MCF7-1

#exp2
reference=LCC2-1
sample=LCC9-1

#exp3
reference=LCC9-1
sample=MCF7-1

#exp4
reference=LCC2-2
sample=MCF7-2

#exp5
reference=LCC2-2
sample=LCC9-2

#exp6
reference=LCC9-2
sample=MCF7-2

# Overlap summit of reference control peaks with sample enriched regions and reference control peaks
TOTAL=$(cat ${reference}_macs_confident.txt | wc -l)
awk -vOFS='\t' '{ $2=$2+$4; $3=$3+1; print $0 }' ${reference}_macs_confident.txt | intersectBed -a stdin -b ${sample}_macs_enrichment.txt | wc -l | awk -vTO TAL=$TOTAL '{ print TOTAL,$1,$1*100/TOTAL }' > cons_conf_enrich_${reference}_${sample}.txt
awk -vOFS='\t' '{ $2=$2+$4; $3=$3+1; print $0 }' ${reference}_macs_confident.txt | intersectBed -a stdin -b ${reference}_macs_control.txt | wc -l | awk -vTO TAL=$TOTAL '{ print TOTAL,$1,$1*100/TOTAL }' > cons_conf_control_${reference}_${sample}.txt
```

– Analyse quantitative

Elle commence d'abord et avant tout par la définition des régions enrichies en étendant, de chaque côté du sommet d'un pic, de la moitié de la longueur moyenne des fragments. On parlera désormais d'une région de pic et non d'un pic. On fusionne par la suite toutes les régions de pics, appelés de manière indépendante dans les différentes paires de conditions, en faisant l'union des coordonnées chromosomiques. Pour chaque région, on assigne le nombre de lectures correspondant à la même position dans le fichier de densité généré précédemment (voir le Calcul de la densité des lectures de la section 4.2.5.4. précédente). On normalise (voir la section suivante) ce nombre par rapport au nombre total des lectures alignées dans chaque échantillon afin d'obtenir un score pour chaque région de pic.

On a utilisé dans cette étape une longueur de région de pic fixe afin de ne pas biaiser l'analyse avec de longues régions de pics. Pour ce faire, on a pris la longueur moyenne des fragments, à savoir une longueur de 250 pb (donc, on étend de 125 pb de chaque côté du sommet des pics).

```
# Define regions with a confident peak in any sample as the region around the peak summit

SIZE=125
pair=MCF7-1;
awk -vOFS='\t' -vSIZE=SIZE '{s=$2+$4-SIZE;e=$2+$4+SIZE;print $1,s,e}' ${pair}_macs_confident.txt
; done | sort -k1,1 -k2,2n | mergeBed -i stdin > peak_regions_1.txt
pair=MCF7-2;
awk -vOFS='\t' -vSIZE=SIZE '{s=$2+$4-SIZE;e=$2+$4+SIZE;print $1,s,e}' ${pair}_macs_confident.txt
; done | sort -k1,1 -k2,2n | mergeBed -i stdin > peak_regions_2.txt
pair=LCC2-1;
awk -vOFS='\t' -vSIZE=SIZE '{s=$2+$4-SIZE;e=$2+$4+SIZE;print $1,s,e}' ${pair}_macs_confident.txt
; done | sort -k1,1 -k2,2n | mergeBed -i stdin > peak_regions_3.txt
pair=LCC2-2;
awk -vOFS='\t' -vSIZE=SIZE '{s=$2+$4-SIZE;e=$2+$4+SIZE;print $1,s,e}' ${pair}_macs_confident.txt
; done | sort -k1,1 -k2,2n | mergeBed -i stdin > peak_regions_4.txt
pair=LCC9-1;
awk -vOFS='\t' -vSIZE=SIZE '{s=$2+$4-SIZE;e=$2+$4+SIZE;print $1,s,e}' ${pair}_macs_confident.txt
; done | sort -k1,1 -k2,2n | mergeBed -i stdin > peak_regions_5.txt
pair=LCC9-2;
awk -vOFS='\t' -vSIZE=SIZE '{s=$2+$4-SIZE;e=$2+$4+SIZE;print $1,s,e}' ${pair}_macs_confident.txt
; done | sort -k1,1 -k2,2n | mergeBed -i stdin > peak_regions_6.txt
```

```
# For each sample and each region add the ratio of chip_read_density / input_read_density

#LCC2-1 vs LCC9-1
cat peak_region_5.txt peak_region_3.txt | sort -k 1,1 -k2,2 | mergeBed -i stdin > peak_regions_LCC2-1_vs_LCC9-1.txt
gunzip -c ${input}.density.gz | intersectBed -a peak_regions_LCC2-1_vs_LCC9-1.txt -b stdin -wao | awk '{peak=$1":"$2":"$3;i
f(old&&peak!=old){print max[old]+0;delete max[old]};if(!max[peak])|max[peak]>$(NF-1)}{max[peak] = $(NF-1)};old = peak}END
{print max[old]+0}' > LCC2-1_vs_LCC9-1_tmp1
paste LCC2-1_vs_LCC9-1_tmp1 LCC2-1_vs_LCC9-1_tmp1 | awk 'if (2==0) {print NA} else {print 1/$2}' | paste peak_regions_LCC
2-1_vs_LCC9-1.txt -> LCC2-1_vs_LCC9-1_tmp ; mv LCC2-1_vs_LCC9-1_tmp peak_regions_LCC2-1_vs_LCC9-1.txt ; rm LCC2-1_vs_LCC9-
1_tmp1

#LCC2-1 vs MCF7-1
cat peak_region_1.txt peak_region_3.txt | sort -k 1,1 -k2,2 | mergeBed -i stdin > peak_regions_LCC2-1_vs_MCF7-1.txt
gunzip -c ${input}.density.gz | intersectBed -a peak_regions_LCC2-1_vs_MCF7-1.txt -b stdin -wao | awk '{peak=$1":"$2":"$3;i
f(old&&peak!=old){print max[old]+0;delete max[old]};if(!max[peak])|max[peak]>$(NF-1)}{max[peak] = $(NF-1)};old = peak}END
{print max[old]+0}' > LCC2-1_vs_MCF7-1_tmp1
paste LCC2-1_vs_MCF7-1_tmp1 LCC2-1_vs_MCF7-1_tmp1 | awk 'if (2==0) {print NA} else {print 1/$2}' | paste peak_regions_LCC
2-1_vs_MCF7-1.txt -> LCC2-1_vs_MCF7-1_tmp ; mv LCC2-1_vs_MCF7-1_tmp peak_regions_LCC2-1_vs_MCF7-1.txt ; rm LCC2-1_vs_MCF7-
1_tmp1

#LCC9-1 vs MCF7-1
cat peak_region_1.txt peak_region_5.txt | sort -k 1,1 -k2,2 | mergeBed -i stdin > peak_regions_LCC9-1_vs_MCF7-1.txt
gunzip -c ${input}.density.gz | intersectBed -a peak_regions_LCC9-1_vs_MCF7-1.txt -b stdin -wao | awk '{peak=$1":"$2":"$3;i
f(old&&peak!=old){print max[old]+0;delete max[old]};if(!max[peak])|max[peak]>$(NF-1)}{max[peak] = $(NF-1)};old = peak}END
{print max[old]+0}' > LCC9-1_vs_MCF7-1_tmp1
paste LCC9-1_vs_MCF7-1_tmp1 LCC9-1_vs_MCF7-1_tmp1 | awk 'if (2==0) {print NA} else {print 1/$2}' | paste peak_regions_LCC
9-1_vs_MCF7-1.txt -> LCC9-1_vs_MCF7-1_tmp ; mv LCC9-1_vs_MCF7-1_tmp peak_regions_LCC9-1_vs_MCF7-1.txt ; rm LCC9-1_vs_MCF7-
1_tmp1

#LCC2-2 vs LCC9-2
cat peak_region_4.txt peak_region_6.txt | sort -k 1,1 -k2,2 | mergeBed -i stdin > peak_regions_LCC2-2_vs_LCC9-2.txt
gunzip -c ${input}.density.gz | intersectBed -a peak_regions_LCC2-2_vs_LCC9-2.txt -b stdin -wao | awk '{peak=$1":"$2":"$3;i
f(old&&peak!=old){print max[old]+0;delete max[old]};if(!max[peak])|max[peak]>$(NF-1)}{max[peak] = $(NF-1)};old = peak}END
{print max[old]+0}' > LCC2-2_vs_LCC9-2_tmp1
paste LCC2-2_vs_LCC9-2_tmp1 LCC2-2_vs_LCC9-2_tmp1 | awk 'if (2==0) {print NA} else {print 1/$2}' | paste peak_regions_LCC
2-2_vs_LCC9-2.txt -> LCC2-2_vs_LCC9-2_tmp ; mv LCC2-2_vs_LCC9-2_tmp peak_regions_LCC2-2_vs_LCC9-2.txt ; rm LCC2-2_vs_LCC9-
2_tmp1

#LCC2-2 vs MCF7-2
cat peak_region_2.txt peak_region_4.txt | sort -k 1,1 -k2,2 | mergeBed -i stdin > peak_regions_LCC2-2_vs_MCF7-2.txt
gunzip -c ${input}.density.gz | intersectBed -a peak_regions_LCC2-2_vs_MCF7-2.txt -b stdin -wao | awk '{peak=$1":"$2":"$3;i
f(old&&peak!=old){print max[old]+0;delete max[old]};if(!max[peak])|max[peak]>$(NF-1)}{max[peak] = $(NF-1)};old = peak}END
{print max[old]+0}' > LCC2-2_vs_MCF7-2_tmp1
paste LCC2-2_vs_MCF7-2_tmp1 LCC2-2_vs_MCF7-2_tmp1 | awk 'if (2==0) {print NA} else {print 1/$2}' | paste peak_regions_LCC
2-2_vs_MCF7-2.txt -> LCC2-2_vs_MCF7-2_tmp ; mv LCC2-2_vs_MCF7-2_tmp peak_regions_LCC2-2_vs_MCF7-2.txt ; rm LCC2-2_vs_MCF7-
2_tmp1

#LCC9-2 vs MCF7-2
cat peak_region_6.txt peak_region_2.txt | sort -k 1,1 -k2,2 | mergeBed -i stdin > peak_regions_LCC9-2_vs_MCF7-2.txt
gunzip -c ${input}.density.gz | intersectBed -a peak_regions_LCC9-2_vs_MCF7-2.txt -b stdin -wao | awk '{peak=$1":"$2":"$3;i
f(old&&peak!=old){print max[old]+0;delete max[old]};if(!max[peak])|max[peak]>$(NF-1)}{max[peak] = $(NF-1)};old = peak}END
{print max[old]+0}' > LCC9-2_vs_MCF7-2_tmp1
paste LCC9-2_vs_MCF7-2_tmp1 LCC9-2_vs_MCF7-2_tmp1 | awk 'if (2==0) {print NA} else {print 1/$2}' | paste peak_regions_LCC
9-2_vs_MCF7-2.txt -> LCC9-2_vs_MCF7-2_tmp ; mv LCC9-2_vs_MCF7-2_tmp peak_regions_LCC9-2_vs_MCF7-2.txt ; rm LCC9-2_vs_MCF7-
2_tmp1
```

- Normalisation

La méthode de normalisation est le facteur clé pour le genre d'analyse dont on dispose. Dans notre cas, comme on compare des conditions au sein d'une même espèce, dans le même type de tissu (tissu mammaire), qu'un même anticorps est utilisé, et que la série d'expériences provient du même laboratoire et très probablement faite côte à côte, cela minimise les différences dans les ratios signal sur bruit (S/N) dus à la variabilité expérimentale. En utilisant cette stratégie, les données

Chip-seq n'ont pas besoin d'être normalisées l'une à l'autre (autre que par le nombre total de reads dans la librairie). Ainsi, on a normalisé les nombres de reads aux tailles des librairies tout en mettant l'hypothèse que le bruit de fond est uniforme entre les conditions. Cela permet les comparaisons de différentes conditions, même si le nombre total et la hauteur des pics est prévu de changer. Cette méthode ne serait pas recommandée lors de l'utilisation de différents anticorps ou différentes espèces. En effet, dans ces cas, les ratios signal sur bruit pourraient ne pas être comparables entre les expériences en raison des différences dans les affinités des anticorps.

En outre, les hauteurs des pics et les emplacements génomiques correspondants ont été normalisés en utilisant la normalisation quantile, une méthode fréquemment utilisée dans l'analyse des données de microarray pour rendre deux distributions identiques.

```
#remove regions with no reads
for pair in LCC2-1_vs_LCC9-1 LCC2-1_vs_MCF7 LCC9-1_vs_MCF7-1 LCC2-2_vs_LCC9-2 LCC2-2_vs_MCF7 LCC9-2_vs_MCF7-2 ; do awk
'($4!=06655!=0)' peak_regions_${pair}.txt > peak_regions_${pair}_no0.txt ; done

>R #Enter
library(preprocessCore) # Load library

table_pre_norm=read.table("peak_regions_ LCC2-1_vs_LCC9-1 _no0.txt") # Load table
table_post_norm=normalize.quantiles(as.matrix(table_pre_norm[,4:5])) # Normalize table
write.table(cbind(table_pre_norm[,1:3],signif(table_post_norm)), "peak_regions_norm_ LCC2-1_vs_LCC9-1.txt", quote=F, sep="\t",
row.names=F, col.names=F) # Save table

table_pre_norm=read.table("peak_regions_ LCC2-1_vs_MCF7-1 _no0.txt") # Load table
table_post_norm=normalize.quantiles(as.matrix(table_pre_norm[,4:5])) # Normalize table
write.table(cbind(table_pre_norm[,1:3],signif(table_post_norm)), "peak_regions_norm_ LCC2-1_vs_MCF7-1.txt", quote=F, sep="\t",
row.names=F, col.names=F) # Save table

table_pre_norm=read.table("peak_regions_ MCF7-1_vs_LCC9-1 _no0.txt") # Load table
table_post_norm=normalize.quantiles(as.matrix(table_pre_norm[,4:5])) # Normalize table
write.table(cbind(table_pre_norm[,1:3],signif(table_post_norm)), "peak_regions_norm_MCF7-1_vs_LCC9-1.txt", quote=F, sep="\t",
row.names=F, col.names=F) # Save table

table_pre_norm=read.table("peak_regions_ LCC2-2_vs_LCC9-2 _no0.txt") # Load table
table_post_norm=normalize.quantiles(as.matrix(table_pre_norm[,4:5])) # Normalize table
write.table(cbind(table_pre_norm[,1:3],signif(table_post_norm)), "peak_regions_norm_ LCC2-2_vs_LCC9-2.txt", quote=F, sep="\t",
row.names=F, col.names=F) # Save table

table_pre_norm=read.table("peak_regions_ LCC2-2_vs_MCF7-2 _no0.txt") # Load table
table_post_norm=normalize.quantiles(as.matrix(table_pre_norm[,4:5])) # Normalize table
write.table(cbind(table_pre_norm[,1:3],signif(table_post_norm)), "peak_regions_norm_ LCC2-2_vs_MCF7-2.txt", quote=F, sep="\t",
row.names=F, col.names=F) # Save table

table_pre_norm=read.table("peak_regions_ MCF7-2_vs_LCC9-2 _no0.txt") # Load table
table_post_norm=normalize.quantiles(as.matrix(table_pre_norm[,4:5])) # Normalize
write.table(cbind(table_pre_norm[,1:3],signif(table_post_norm)), "peak_regions_norm_MCF7-2_vs_LCC9-2.txt", quote=F, sep=
"\t", row.names=F, col.names=F) # Save table

> q()
> n
```

- Changements quantitatifs

Cette étape détermine les changements quantitatifs en calculant les différences entre les hauteurs de pics en terme de “fold change” log2. On assigne les différentes régions à une catégorie des changements quantitatifs sur la base de la variation de la densité des lectures normalisée :

- Catégorie invariants ($-2 \text{ fold} < \text{score} < 2 \text{ fold}$).
- Catégorie croissants ($\text{score} > 2 \text{ fold}$).
- Catégorie décroissants ($\text{score} < -2 \text{ fold}$).

```
# Calculate log2(change)
for pair in LCC2-1_vs_LCC9-1 LCC2-1_vs_MCF71 LCC9-1_vs_MCF7-1 LCC2-2_vs_LCC9-2 LCC2-2_vs_MCF72 LCC9-2_vs_MCF7-2 ; do

grep -v "NA" peak_regions_norm_${pair}.txt | awk -v OFS="\t" '{print $0, log($4/$5)/log(2)}' > peak_regions_norm_${pair}_log2.txt

# Regions 2 fold higher in conditionA than conditionB
awk '($6>=2)' peak_regions_norm_${pair}_log2.txt > peak_regions_norm_${pair}_log2_decrease.txt

# Regions with no quantitative changes (within 2 fold)
awk '($6>=2&&$6<2)' peak_regions_norm_${pair}_log2.txt > peak_regions_norm_log2_invariant_${pair}.txt

# Regions 2 fold lower in conditionA than conditionB
awk '($6<=-2)' peak_regions_norm_${pair}_log2.txt > peak_regions_norm_log2_increase_${pair}.txt

# Count number of regions
wc -l peak_regions_norm_${pair}_log2_*.txt

done
```

4.2.5.5. Étape d'analyse tertiaire

- Annotation des locations des pics

À la suite du “peak calling” et de la catégorisation des pics, on procède à l'annotation des régions des pics en utilisant le pipeline de génome Québec. Plusieurs métriques

sont générées, telles que la moyenne des largeurs des pics, la moyenne des hauteurs des pics et le pourcentage de pics à proximité des sites de transcription (-1000, 1000 pb). En plus, les pics sont annotés par rapport à leur localisation dans le génome :

- Gène (exon ou intron)
- Proximal (0,2] kb en amont du début d'un site de transcription.
- Distal (2,10] kb en amont du début d'un site de transcription.
- 5d (10,100] kb en amont du début d'un site de transcription.
- Désert de gène (Gene desert) ≥ 100 kb en amont ou en aval du début d'un site de transcription.
- Autre (pas inclus dans les catégories précédentes)

– Analyse des motifs

Pour l'analyse des motifs, le pipeline mugqic utilise l'outil HOMER. Ce dernier est une implémentation d'un algorithme de découverte de motifs qui a été conçu pour l'analyse des éléments régulateurs en génomique. Il s'agit d'un algorithme de reconnaissance de motif différentiel, ce qui signifie qu'il prend deux ensembles de séquences et tente d'identifier les éléments de régulation qui sont spécifiquement enrichis dans l'un par rapport à l'autre. Il utilise le score ZOOPS (zéro ou une occurrence par séquence) couplé avec les calculs d'enrichissement hypergéométriques (ou binomiaux) pour déterminer l'enrichissement de motif. HOMER tient compte des différents biais de séquençage dans le jeu de données tel que le biais GC (pour plus de détails se référer à la section 'Analyse des séquences', où, une description détaillée de HOMER est présentée).

```
[aouza123@lg-1r17-n04 chipseq]$ for pair in LCC2-1_vs_LCC9-1 LCC2-1_vs_MCF71 LCC9-1_vs_MCF7-1 LCC2-2_vs_LCC9-2 LCC2-2_vs_MCF72 LCC9-2_vs_MCF7-2 ; do mkdir decrease_${pair}/peak_call/Narrow/ ; mkdir increase_${pair}/peak_call/Narrow/ ; mkdir invariant_${pair}/peak_call/Narrow/ ; cp ../13/Quantitative_changes/peak_regions_norm_log2_decrease_${pair}.txt decrease_${pair}/peak_call/Narrow/Narrow_peaks.narrowPeak ; cp ../13/Quantitative_changes/peak_regions_norm_log2_increase_${pair}.txt increase_${pair}/peak_call/Narrow/Narrow_peaks.narrowPeak ; cp ../13/Quantitative_changes/peak_regions_norm_log2_invariant_${pair}.txt invariant_${pair}/peak_call/Narrow/Narrow_peaks.narrowPeak ; done

[aouza123@lg-1r17-n04 chipseq]$ for pair in LCC2-1_vs_LCC9-1 LCC2-1_vs_MCF71 LCC9-1_vs_MCF7-1 LCC2-2_vs_LCC9-2 LCC2-2_vs_MCF72 LCC9-2_vs_MCF7-2 ; do cp decrease_LCC2-1_vs_LCC9-1/design.csv decrease_${pair}/ ; cp decrease_LCC2-1_vs_LCC9-1/project.nanuq.csv decrease_${pair}/ ; cp decrease_LCC2-1_vs_LCC9-1/chipseq.base.ini decrease_${pair}/ ; cp decrease_LCC2-1_vs_LCC9-1/design.csv increase_${pair}/ ; cp decrease_LCC2-1_vs_LCC9-1/project.nanuq.csv increase_${pair}/ ; cp decrease_LCC2-1_vs_LCC9-1/chipseq.base.ini increase_${pair}/ ; cp decrease_LCC2-1_vs_LCC9-1/design.csv invariant_${pair}/ ; cp decrease_LCC2-1_vs_LCC9-1/project.nanuq.csv invariant_${pair}/ ; cp decrease_LCC2-1_vs_LCC9-1/chipseq.base.ini invariant_${pair}/ ; done

[aouza123@lg-1r17-n04 chipseq]$ for i in `ls -d *vs*` ; do cd $i ; python ../../test_python/mugqic_pipelines/chipseq/chipseq.py -c chipseq.base.ini -s 13-15 -d design.csv -r project.nanuq.csv > to_run.sh ; cd .. ; done
```

```
#-----
# STEP: homer_find_motifs_genome
#-----
STEP=homer_find_motifs_genome
mkdir -p $JOB_OUTPUT_DIR/$STEP

#-----
# JOB: homer_find_motifs_genome_1_JOB_ID: homer_find_motifs_genome.Narrow
#-----
JOB_NAME=homer_find_motifs_genome.Narrow
JOB_DEPENDENCIES=$macs2_callpeak_1_JOB_ID
JOB_DONE=job_output/homer_find_motifs_genome/homer_find_motifs_genome.Narrow.72bc52a115d101cf53bc9ede7877433d.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_${TIMESTAMP}.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'homer_find_motifs_genome.Narrow.72bc52a115d101cf53bc9ede7877433d.mugqic.done'
module load mugqic/perl/5.18.2 mugqic/weblogo/2.8.2 mugqic/homer/4.7 && \
mkdir -p annotation/Narrow/Narrow && \
findMotifsGenome.pl \
    peak_call/Narrow/Narrow_peaks.narrowPeak \
    hg19 \
    annotation/Narrow/Narrow \
    -preparedDir annotation/Narrow/Narrow/prepared \
    -p 4
homer_find_motifs_genome.Narrow.72bc52a115d101cf53bc9ede7877433d.mugqic.done
)
homer_find_motifs_genome_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND"
MUGQIC_STATE=\$PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ] ; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:00 -q sw -l nodes=1:ppn=4 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "homer_find_motifs_genome_1_JOB_ID          $JOB_NAME          $JOB_DEPENDENCIES          $JOB_OUTPUT_RELATIVE_PATH" >> $JOB_LIST
```

– Étude de la reproductibilité entre répliquats

Une méthode puissante pour évaluer la similitude globale, entre deux modèles de liaison de facteurs de transcription, est le PCC entre les densités de reads respectives à l'échelle du génome (le nombre de reads à chaque position dans le génome).

Comme le PCC est un seuil indépendant, il élimine certains défis associés à l'utilisation des seuils pour l'appel de pics, et est plus robuste face aux variations expérimentales des hauteurs de pics. Ainsi, en utilisant les fichiers de pics comme entrée et le script `correlation.awk`, on a obtenu les PCCs (voir les diagrammes de Venn dans la partie résultats) pour évaluer la similitude entre nos réplicats biologiques dans nos différentes conditions.

- Annotation des pics identifiés

Les pics, identifiés et classés par catégories (changements quantitatifs), sont annotés en utilisant Gene ontology (GO) et genome ontology intégrés dans le pipeline muggic.

```

#-----
# STEP: homer_annotate_peaks
#-----
STEP=homer_annotate_peaks
mkdir -p $JOB_OUTPUT_DIR/$STEP

#-----
# JOB: homer_annotate_peaks_1_JOB_ID: homer_annotate_peaks.Narrow
#-----
JOB_NAME=homer_annotate_peaks.Narrow
JOB_DEPENDENCIES=$macs2_callpeak_1_JOB_ID
JOB_DONE=job_output/homer_annotate_peaks/homer_annotate_peaks.Narrow.acf7db9dd15de3c0241a36315af41d82.mugqic.done
JOB_OUTPUT_RELATIVE_PATH=$STEP/${JOB_NAME}_${TIMESTAMP}.o
JOB_OUTPUT=$JOB_OUTPUT_DIR/$JOB_OUTPUT_RELATIVE_PATH
COMMAND=$(cat << 'homer_annotate_peaks.Narrow.acf7db9dd15de3c0241a36315af41d82.mugqic.done'
module load mugqic/perl/5.18.2 mugqic/homer/4.7 mugqic/mugqic_tools/2.0.3 && \
mkdir -p annotation/Narrow/Narrow && \
annotatePeaks.pl \
  peak_call/Narrow/Narrow_peaks.narrowPeak \
  hg19 \
  -gsize hg19 \
  -cons -CpG \
  -go annotation/Narrow/Narrow \
  -genomeOntology annotation/Narrow/Narrow \
  > annotation/Narrow/Narrow.annotated.csv && \
perl -MReadMetrics -e 'ReadMetrics::parseHomerAnnotations(
  "annotation/Narrow/Narrow.annotated.csv",
  "annotation/Narrow/Narrow",
  -2000,
  -10000,
  -10000,
  -100000,
  100000
)'
homer_annotate_peaks.Narrow.acf7db9dd15de3c0241a36315af41d82.mugqic.done
)
homer_annotate_peaks_1_JOB_ID=$(echo "rm -f $JOB_DONE && $COMMAND
MUGQIC_STATE=\$PIPESTATUS
echo MUGQICexitStatus:\$MUGQIC_STATE
if [ \$MUGQIC_STATE -eq 0 ] ; then touch $JOB_DONE ; fi
exit \$MUGQIC_STATE" | \
qsub -m ae -M $JOB_MAIL -W umask=0002 -d $OUTPUT_DIR -j oe -o $JOB_OUTPUT -N $JOB_NAME -l walltime=24:00:
0 -q sw -l nodes=1:ppn=1 -W depend=afterok:$JOB_DEPENDENCIES | grep "[0-9]")
echo "$homer_annotate_peaks_1_JOB_ID $JOB_NAME $JOB_DEPENDENCIES $JOB_OUTPUT_RELATIVE_PATH
" >> $JOB_LIST

```

Le tableau 11, ci dessous, donne un aperçu global des principaux scripts sus-cités en décrivant leur tâche et leur fichiers en entrée (input) et en sortie (output). Il est important de noter qu'un output de type métrique est soit des statistiques soit des données interprétables telles que les noms ou les locations des motifs.

Tableau 11 : Récapulatif des principaux scripts décrivant les étapes d'analyse.

Script		Tâche	Input	Output
clean_bam.sh		Suppression des alignements secondaires	Fichier bam	Fichier bam filtré
bam_to_fastq.sh		Conversion des bams en fastq	Fichier bam	Fichier Fastq
guess_phred.sh		Encodage de qualité	Fastq	Métrique
manual_phred.sh		Encodage de qualité	Fastq	Métrique
samtools view		Filtrations des reads alignées	Fichier bam	Fichier bam filtré
ChIPseq.py	Trimmomatic	Trimming de qualité et clipping des adaptateurs	Fichiers Fastq	Fichier Fastq trimmé
	Bwa	Alignement	Fichier Fastq trimmé	Fichier bam
	picard/SortSam.jar	Tri des fichiers bam	Fichier bam	Fichier Fastq trié
	Samtools index	Indexage des fichiers bam	Fichier bam	index bai
	Samtools view	Filtration des Bams	Fichier bam	Fichier bam filtré
	picard/MergeSam	Fusion des bams	Fichiers bam	Fichier bam
	picard/MarkDuplicate	Marquage des doublons	Fichier bam	Fichiers bam sans doublons
	MACS	Calling des pics	Fichier bam	fichier bed (Pics)
Quant_analysis.sh		Analyse quantitative et catégorisation des pics	fichier bed (Pics)	Métrique
ChIPseq.py	HOMER	Annotation des locations des pics	fichier bed (Pics)	Métrique
	Command beach	Filtration des pics et création des pics de contrôle	fichier bed (Pics)	fichier bed (Pics filtrés)
	HOMER	Analyses des motifs	fichier bed (Pics)	Métriques
	HOMER	Annotation des pics identifiés	fichier bed (Pics)	Métrique

CHAPITRE V

RÉSULTATS ET DISCUSSION

Afin de comparer les profils des sites de liaison d'un facteur de transcription entre plusieurs conditions biologiques, une méthode bio-informatique a été mise en place. Elle consiste à combiner des étapes du pipeline d'analyse ChIP-Seq de génome québécois avec une partie d'analyse comparative intégrée et une phase préalable de traitement des données. Cette dernière traite le cas de séquences déjà alignées sur des versions plus anciennes du génome cible. L'analyse comparative, quant à elle, filtre et catégorise les sites de liaison en deux étapes indépendantes, une méthode complète peu employée dans la littérature. La première caractérise le nombre de sites de liaison spécifiques à chaque condition ainsi que le nombre de sites de liaison partagés entre les conditions deux à deux. De son côté, la deuxième étape applique une analyse quantitative, basée sur le calcul de la densité des reads au niveau des sites de liaisons, afin d'en extraire les liaisons différentielles.

L'identification et l'analyse des sites de liaison avec une expérience ChIP-Seq peuvent être effectuées de plusieurs façons en fonction des logiciels et des outils choisis pour agencer les différentes étapes du pipeline. Ainsi, différents séquenceurs, aligneurs, "peak callers", programmes de recherche des motifs sur-représentés et algorithmes d'identification des motifs peuvent être utilisés. Cependant, tous les programmes disponibles au niveau de chaque étape du pipeline, avec chacun ses avantages et ses inconvénients, ne donnent pas les mêmes résultats. Une étude comparative des résultats et de l'efficacité de tous ces programmes serait intéressante, mais ceci ne

rentre pas dans le cadre du projet. En outre, dans l'ordre mentionné, ces logiciels forment des pipelines qui n'offrent que la possibilité d'une étude des sites de liaison d'une expérience Chip-seq singulière, tel est le cas du pipeline de génome québec.

Notre ensemble d'analyse a été appliqué pour identifier et comparer le répertoire des sites de liaison du facteur de transcription ERa, le cistrome ERa, dans trois lignées cellulaires du cancer du sein. Sachant que la stimulation par l'estrogène mène à l'établissement de programme de transcription spécifique dans les cellules du cancer du sein ERa-positif, l'objectif était de comparer ce programme de transcription entre les cellules du cancer du sein sensibles (la lignée MCF-7) et les cellules du cancer du sein résistantes (la lignée LCC-2 au tamoxifène et la lignée LCC-9 au tamoxifène et au fulvestrant) à l'hormonothérapie, afin de mieux comprendre les mécanismes moléculaires liés à l'acquisition de la résistance aux médicaments dans le traitement du cancer du sein.

Prétraitement des données

Les données de départ étaient des fichiers Bams contenant des alignements de lectures sur le génome humain hg18. Afin d'avoir l'alignement sur la dernière version du génome humain hg19, nous avons d'abord converti nos fichiers Bam en fichiers Fastq. Avec cette conversion, nous risquons d'avoir plusieurs versions d'une même lecture (les doublons) sans que celles-ci aient été générées initialement par le séquenceur. Ceci est principalement dû au fait qu'il peut y avoir plusieurs alignements pour la même séquence. Ainsi, nous avons filtré nos six bam en exploitant quelques caractéristiques du format sam/bam. Ce format distingue entre les alignements primaires et secondaires par des "flag" différents en faisant appel à une instance du logiciel samtools (Li, 2009) (qui permet de lire le fichier bam sous forme binaire). Le nombre de lectures obtenues après la conversion des Bam en Fastq dans les trois lignées cellulaires est présenté dans le tableau 12. Notez que du fait de la nature des

données de départ (données déjà alignées), le nombre total de lectures brutes ne correspond pas au nombre généré par le séquenceur, mais c'est plutôt le nombre de lectures, filtrées selon la qualité du premier alignement, sur le génome hg18 et après suppression des doublons.

En outre, l'analyse en aval des données demande une connaissance préalable de la nature des séquences ainsi que la technologie de séquençage avec laquelle ils ont été générés. Il est aussi important de connaître la taille des séquences et leur encodage de qualité. Pour ce faire, afin de déterminer la nature du séquençage utilisé, nous avons calculé les statistiques d'alignement avec samtools flagstat. Les résultats montrent qu'il n'y a aucun read qui est "paired" avec un autre, ce qui nous a amené à conclure que c'est du "single-end". De plus, avant de procéder à l'alignement contre la nouvelle version du génome humain hg19, nous avons investigué en profondeur les fichiers fastq pour mieux déterminer l'origine des séquences et avoir une idée sur leur qualité. Comme nos données étaient relativement anciennes (alignées contre la version hg18 du génome qui date d'avant 2011), nous avons eu besoin de connaître l'encodage pour bien pouvoir interpréter leur qualité. L'analyse des séquences de qualité a montré que c'est du Illumina miseq ou hiseq phred 64 ou bien du Solexa/Illumina phred 64. Compte tenu que l'alignement a été fait versus la version hg18 du génome et que la technologie miseq est très récente, il s'agirait donc de données générées par un séquenceur Illumina/Solexa ou illumina hiseq avec un encodage phred 64. L'analyse a aussi montré que la taille des reads est toujours constante à 40 pb. Cette situation est peu commune vue que le trimming des adaptateurs ne donne pas une distribution de taille aussi uniforme dans les fichiers fastq de sortie. En effet, on aurait dû avoir des reads de tailles différentes (entre 40 pb et 50 pb dans le cas où l'adaptateur serait de longueur 10 pb).

Tableau 12 : Statistiques de l'alignement des lectures et le nombre de pics (FDR 1%) identifiés dans chacune des trois lignées cellulaires du cancer du sein et leur repliquats.

Echantillon	Lectures brutes	Lectures alignées	Taux duplication	Total des pics (FDR 1%)
LCC2-1	9887873	100%	45,29%	18461
LCC2-2	11034748	100%	41,13%	16295
LCC9-1	17548772	100%	43,36%	22253
LCC9-2	8303371	100%	48,26%	14090
MCF7-1	13570499	100%	46,66%	14585
MCF7-2	11716089	100%	44,25%	11783

Comme les données de départ étaient des lectures déjà alignées, cela sous entend qu'elles ont préalablement subies l'étape de la filtration selon la qualité (Quality trimming) ainsi que la suppression des adaptateurs (adapter clipping). Par conséquent, nous sommes directement passés à la prochaine étape du pipeline ChIP-Seq, à savoir l'alignement.

Alignement des reads

Le pipeline mugqc utilise la variante MEM de l'outil d'alignement bwa (Li, 2009) pour aligner les lectures contre la version hg19 du génome humain. Cette variante MEM est désignée pour aligner des séquences dont la taille est supérieure à 100 pb, mais nos séquences étaient de taille 40 pb. Nous n'avons pas changé d'outil d'alignement mais ceci pourrait être envisagé dans une tentative d'optimisation de l'alignement pour des analyses ultérieures.

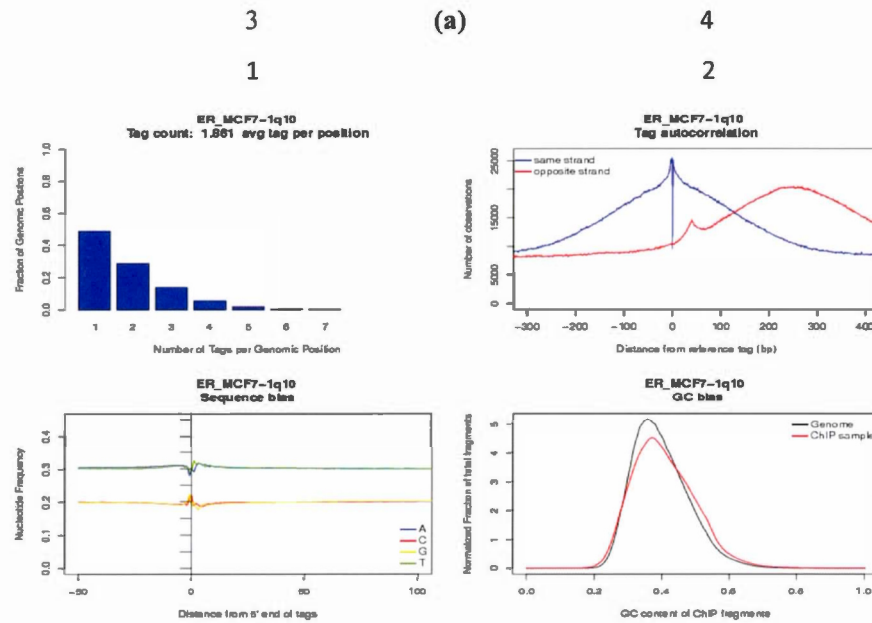
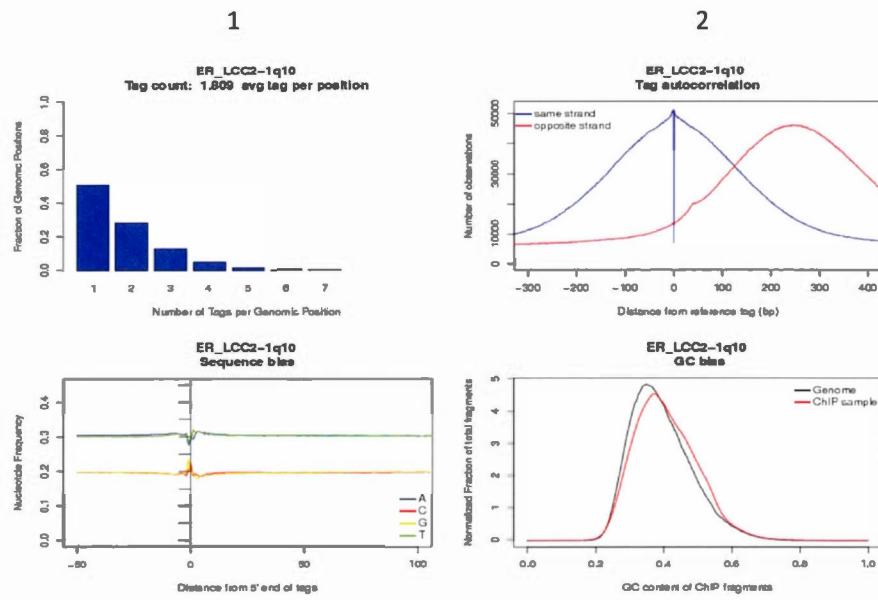
Les résultats d'alignement étaient des fichiers bams qui ont nécessité une série de modifications dans le but de les rendre utilisables par les outils en aval. Comme la plupart des outils requièrent des index .bai pour pouvoir lire ce genre de fichiers, la première modification était leur indexation (en utilisant l'outil samtools). Ensuite, les reads ont été triés par ordre numérique selon leurs positions avec l'outil picard afin de faciliter l'accès. Nous avons aussi supprimé, des fichiers bams, les reads alignés avec une valeur de qualité inférieure au seuil de qualité déterminé (mapq=0, dans notre cas).

En raison d'erreurs inhérentes à la technologie de séquençage, certaines lectures auront des copies exactes issues du même fragment d'ADN. Les doublons partagent la même séquence et la même position d'alignement et pourraient causer des problèmes au cours du peak calling en sur-représentant certaines régions. En plus, pour les approches Single-end la probabilité de les avoir est encore plus importante. Ainsi, Picard a été utilisé pour marquer ces doublons dans le fichier BAM. Ce dernier les définit en leur attribuant un "flag" spécifique. Les outils en aval sont ainsi capables de reconnaître ce "flag" et n'utiliseront pas les lectures marquées. Ce nombre de lectures redondantes, mappées aux mêmes coordonnées génomiques, est une mesure usuelle de contrôle de la qualité. En effet, un taux élevé de redondance suggère un biais d'une amplification PCR à partir de matériel ChIP limité. Cette fraction de reads dupliqués n'est pas à considérer parmi les reads ayant réussi le passage de la filtration par mapq. Le ratio de reads redondants sur tous les reads mappés devrait idéalement être sous 50%. Nos résultats (voir Tableau 12) affichent des taux de duplication en dessous du seuil pour chaque expérience. La moyenne étant de 44.82 % entre les lignées y compris les répliques. Ainsi, après l'alignement des lectures avec BWA, des 13570499, 9887873 et 17548772 lectures alignées dans les cellules MCF7-1, LCC2-1 et LCC9-1, respectivement, seuls 54,71 %, 56,46 % et 53,34 % sont alignées de façon unique. De la même façon, dans les répliques MCF7-2,

LCC2-2 et LCC9-2, seuls 55,75%, 58,87 %, et 51,74 % des 11716089, 11034748 et 8303371 lectures respectivement, sont uniques.

Pour faciliter l'analyse ChIP-Seq, le pipeline mugqic intègre le logiciel HOMER4 (Heinz, 2010). Ce dernier transforme l'alignement de séquence en une structure de données indépendante de la plate-forme. Toutes les informations pertinentes à propos de l'expérience sont organisées en un répertoire de "Tag". Plusieurs fonctions de contrôle de qualité sont exécutées pour aider à fournir des informations et des commentaires sur la qualité de l'expérience. Plusieurs paramètres importants sont des estimations qui peuvent être exploitées dans les analyses en aval, tels que la longueur approximative des fragments ChIP-Seq (voir le Tableau S1 en annexe).

Les quatre mesures de qualité obtenues à l'aide du logiciel HOMER sont le nombre de lectures par position unique, le biais de séquence, le biais GC et l'auto-correlation des "tags". Ci dessous, les figures des résultats obtenus pour ces mesures dans chaque expérience (voir Figure 30). Les résultats pour nos expériences replicats sont présentés sur la Figure S1 dans l'annexe.



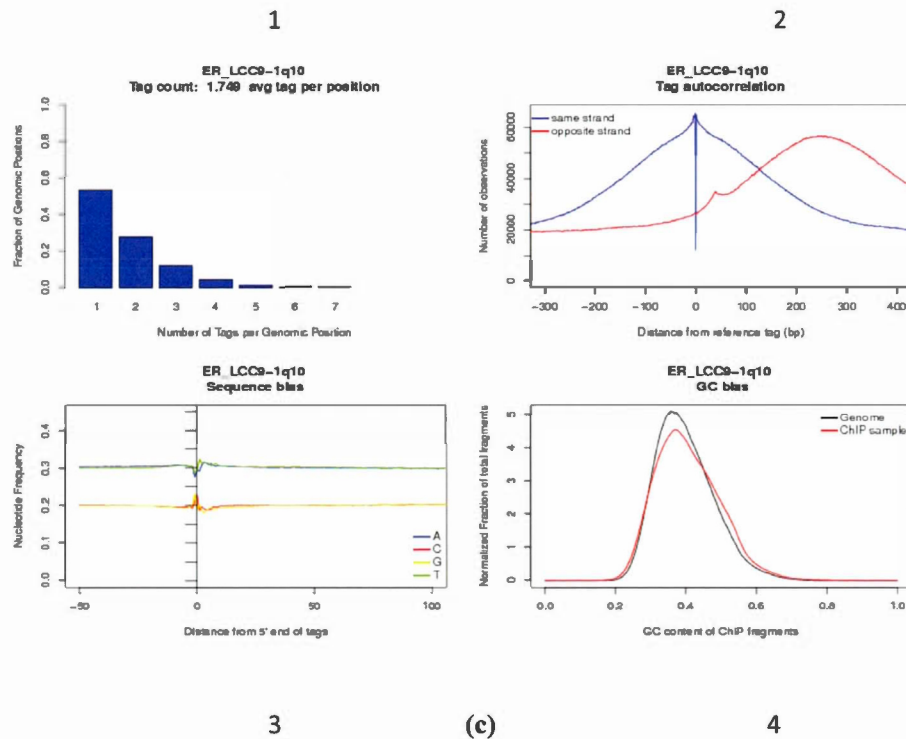


Figure 30 : Les quatre mesures de qualité (le nombre de lecture par position unique, l'auto-corrélation des "tags" le biais de séquence et le biais GC) obtenues à l'aide du logiciel HOMER dans chacune de nos trois expériences ChIP-Seq (a), (b) et (c): (a)1, (b)1 et (c)1 indiquent le nombre de tag par position unique dans chacune des expériences. Si une expérience est sur-séquencée, le nombre de lectures dupliquées devient trop important par rapport au nombre de lectures uniques. Parfois, cela est un signe qu'il n'y avait pas assez de matériel de départ pour la préparation de la librairie de séquençage. Dans ce cas, le nombre moyen de tag par position augmente proportionnellement. Les chiffres 1.80 (a)1, 1.86 (b)1 et 1.74 (c)1 restent acceptables, le nombre de tag moyen < 1.5 étant le seuil généralement recommandé. (a)2, (b)2 et (c)2 L'auto-corrélation crée une distribution des distances entre les lectures adjacentes dans le génome, une sur le brin sens et l'autre sur le brin anti-sens. L'analyse d'auto-corrélation est un bon indicateur de la qualité de l'expérience. (a)3, (b)3 et (c)3 Le biais de séquence montre la relation entre le read séquencé et la séquence génomique réelle. De petites variations dans le début du read peuvent se produire mais dans l'ensemble, les ratios de nucléotides doivent rester constants tel que montré. (a)4, (b)4, et (c)4 Le biais GC montre le contenu en GC pour chaque lecture. Les résultats dans nos trois expériences ne montrent aucun décalage. Si la distribution de GC du read est décalée par rapport à la distribution habituelle du génome, alors le read pourrait être surreprésenté dans les régions riches ou pauvres en GC.

– L’auto-correlation des “tag”

L’analyse de l’auto-corrélation est un bon indicateur de la qualité de l’expérience ChIP-Seq. Elle donne une première idée du déroulement général de ChIP-seq et indique si l’expérience a bien fonctionné ou pas. Elle crée une distribution de la distance entre les lectures adjacentes dans le génome. Deux distributions sont présentées, une pour le brin principal et une pour le brin opposé. Les résultats de l’analyse d’autocorrélation sont très utiles pour résoudre les problèmes avec l’expérience, et sont utilisés pour estimer la longueur des fragments. Nos résultats d’autocorrélation sont présentés dans la Figure 30 (a)2, (b)2 et (c)2. Cette dernière montre qu’on a eu, dans toutes les expériences, des zones où des lectures se sont condensées, témoins de régions enrichies. Les courbes de l’autocorrélation présentent bien deux distributions, une sur le brin positif et l’autre sur le brin négatif. En plus, l’allure des courbes est ni plate ni extrêmement épineuse, ce qui est un signe que les expériences se sont bien déroulées. En effet, la hauteur des courbes (pics) est un déterminant utile qui témoigne de la réussite d’une expérience. Des courbes plates sont un signe de faible signal ou de sites de fixation associés à certains histones qui s’étendent sur de larges régions. Par contre, des courbes trop épineuses témoignent d’une faible complexité des librairies se caractérisant par l’empilement des reads alignés. L’autocorrélation peut aussi nous fournir une estimation de la longueur du fragment utilisée pour le séquençage. Cette estimation correspond au maximum du signal de l’autocorrélation dans les lectures du brin opposé (qui est, d’après les figures d’autocorrélation, aux alentours de 250 pb).

- Le biais GC

La figure correspondant à ce biais (voir Figure 30 (a)4, b(4) et (c)4) montre le contenu en GC pour chaque lecture. Dans toutes les expériences, nos résultats ne montrent aucun décalage. En effet, si la distribution de GC de la lecture était décalée par rapport à la distribution habituelle du génome, la lecture pourrait être surreprésentée dans les régions riches ou pauvres en GC. Il est très facile pour la moyenne de la teneur en GC de la librairie de séquençage de changer, ceci en raison des nombreuses étapes de préparation de la librairie, impliquant l'amplification, la sélection de la taille et de l'extraction au gel. Cela peut être problématique pour les analyses en aval. HOMER vérifie le biais dans les données de séquençage en alignant chaque lecture dans son contexte génomique et en calculant la fréquence de nucléotides à chaque position par rapport à l'extrémité 5' de la lecture. Cela peut être utile pour identifier les biais dans les séquences produites, les problèmes de séquençage, ou des problèmes avec le séquençage préférentiel des régions riches en GC contre celles riches en AT. Le problème avec un échantillon de % GC modifié est que même si l'échantillon est aléatoire, on pourrait avoir un "enrichissement" à des endroits à forte teneur en GC dans le génome, comme au niveau des îlots CpG. C'est dommage, car la plupart des régions riches en GC sont les sites de démarrage de la transcription, ce qui pourrait faire croire que l'expérience est un succès.

- le biais de séquence

L'autre mesure de la qualité est le biais de séquence. Ce dernier montre la relation entre le read séquencé et la séquence génomique réelle. De petites variations dans le début du read peuvent se produire, mais dans l'ensemble, les ratios de nucléotides doivent rester constants, tel que montré dans nos résultats d'expériences (voir Figure 30 (a)3, (b)3 et (c)3).

- Le nombre de lectures par position unique

Le nombre de lectures par position unique est aussi représenté sur la Figure 30 (a)1, (b)1 et (c)1 pour chacune des expériences. Si une expérience contient un nombre de lectures dupliquées trop important, le nombre moyen de “tag” augmente proportionnellement. Nos résultats affichent un nombre de lectures par position unique de 1,80, 1,86 et 1,74 pour ER_LCC2-1, ER_MCF7-1 et ER_LCC9-1, respectivement. Ces chiffres restent acceptables, toutefois, un nombre de tag moyen inférieur à 1,5 est généralement recommandé.

Identification des sites de liaison de ERa

Après l'alignement et afin d'identifier les sites de liaison de ERa, nous avons défini les enrichissements en lectures (pics) dans chacune des expériences indépendamment. Les régions enrichies en lectures dans les trois conditions ont été identifiées en utilisant l'algorithme MACS (voir Tableau S1 dans l'annexe: exemple de fichier de sortie de MACS). Un total de 14 586, 18 462 et 22 254 sites de liaison de haute confiance (FDR 1%) pour ERa ont été trouvés dans les cellules MCF7-1, LCC2-1 et LCC9-1, respectivement (voir Tableau 12). Afin de tester la qualité des sites de liaison trouvés, deux replicats biologiques ont été utilisées pour chaque condition. Ainsi un total de 11 784, 16 296 et 14 091 sites ont été identifiés dans MCF7-2, LCC2-2 et LCC9-2 respectivement. Intuitivement, les enrichissements ChIP réussis devraient être cohérents entre les réplicas biologiques en présentant des profils de signaux et des régions de pics semblables. L'une des mesures de cohérence utilisée, qui peut être facilement représentée par un diagramme de Venn (voir Figure 31), est le pourcentage (> 50%) de pics se chevauchant entre deux réplicas. L'analyse comparative a ainsi montré un taux de 56.94 %, 59.24 % et 49.79 % sites de liaison

en commun entre les deux répliques dans chacune des trois conditions (voir Figure 31). Ceci répond au seuil de contrôle de la qualité communément exigé, qui doit être supérieur à 50%. Nous avons aussi mesuré la reproductibilité des reads en calculant le coefficient de corrélation de Pearson (PCC) du nombre de reads, alignés à chaque position génomique, au niveau de l'union des régions de pics des répliques. L'intervalle de PCC est typiquement de 0.3–0.4 (pour des échantillons sans corrélation - indépendants-) à 0.9 (pour des échantillons replicas dans des expériences de haute qualité). Des valeurs basses suggèrent que l'une ou les deux répliques sont de mauvaise qualité. Nos résultats des mesures sont $PCC = 0.734331$ pour MCF7, $PCC = 0.799899$ pour LCC2 et $PCC = 0.752862$ pour LCC9. Ces mesures sont au dessus du seuil minimal de 0.6, généralement considéré (dans ce genre d'analyse) comme seuil de bonne corrélation entre les répliques.

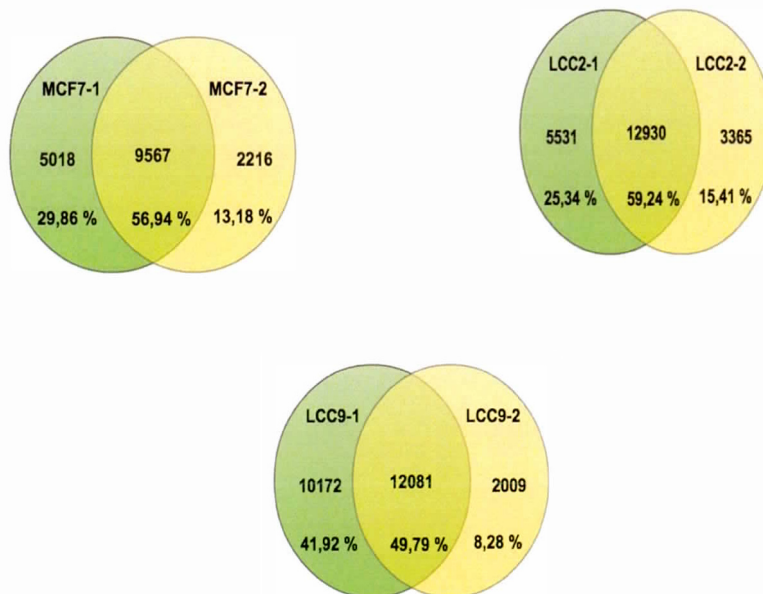
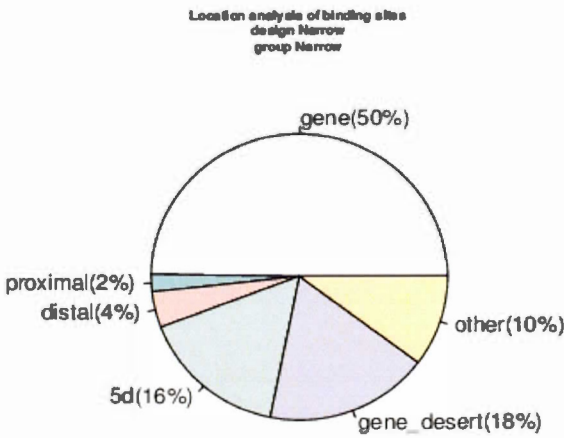


Figure 31: Les diagrammes de Venn des ensembles de pics identifiés à partir des deux répliques individuelles au même cut-off pour le peak calling dans les cellules MCF7, LCC2 et LCC9. Une large intersection indique une haute cohérence entre les répliques (> 50%).

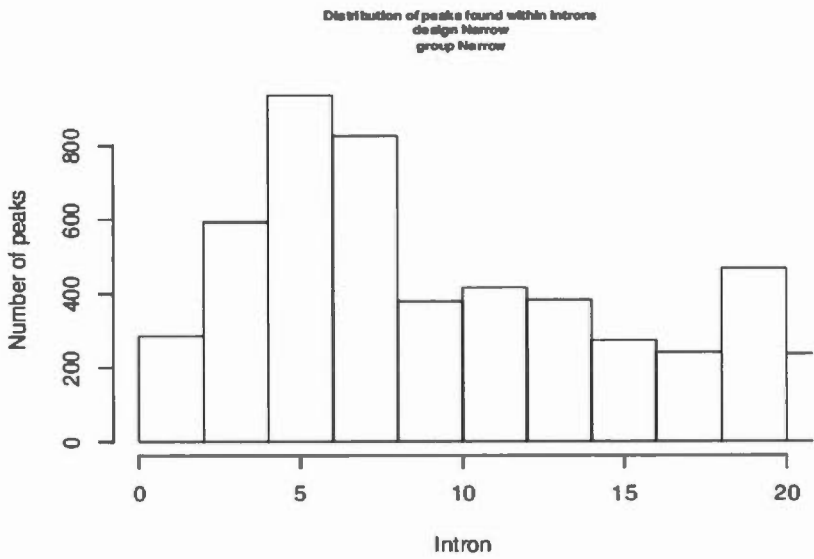
Annotation des localisations des pics

En plus, nous avons examiné la localisation des sites d'enrichissement ERa par rapport aux gènes voisins les plus proches. Les résultats sont présentés dans la Figure 32. Seule une petite fraction des pics, qui varie de 6 à 8%, selon l'expérience, se localise dans les régions promotrices (Proximal, (0,2] kb en amont du début d'un site de transcription, et Distal, (2,10] kb en amont du début d'un site de transcription). Aussi, la plus grande fraction des pics, variant de 48 à 51 % entre les expériences, réside dans les sites intragéniques, dont la majorité dans les introns (voir Figure 32 (a)1, (b)1 et c(1)), ce qui est en accord avec les données publiées (Carroll et al, 2006; Lin et al, 2007). En effet, ces dernières attribuent une action à longue portée loin des régions promotrices pour le facteur de transcription ERa. Les mêmes résultats ont été observés avec les expériences réplicats (voir Figure S3 dans l'annexe).

-1-

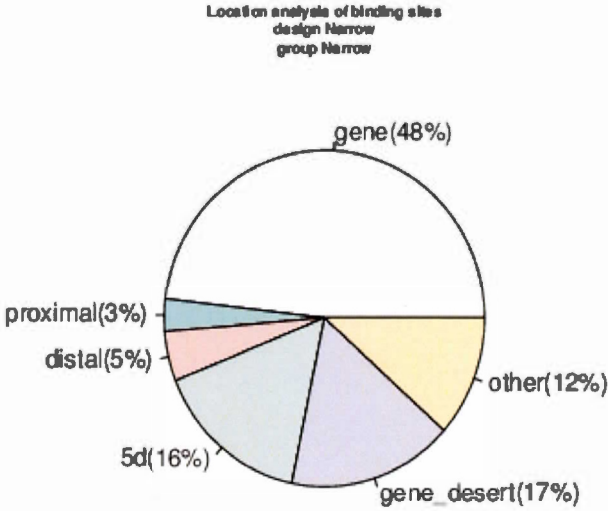


-2-

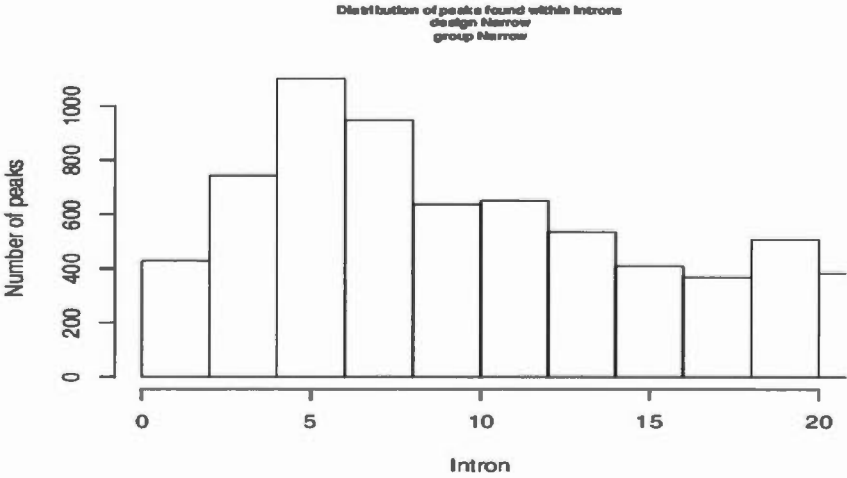


(a): MCF7

-1-



-2-



(b): LCC2-1

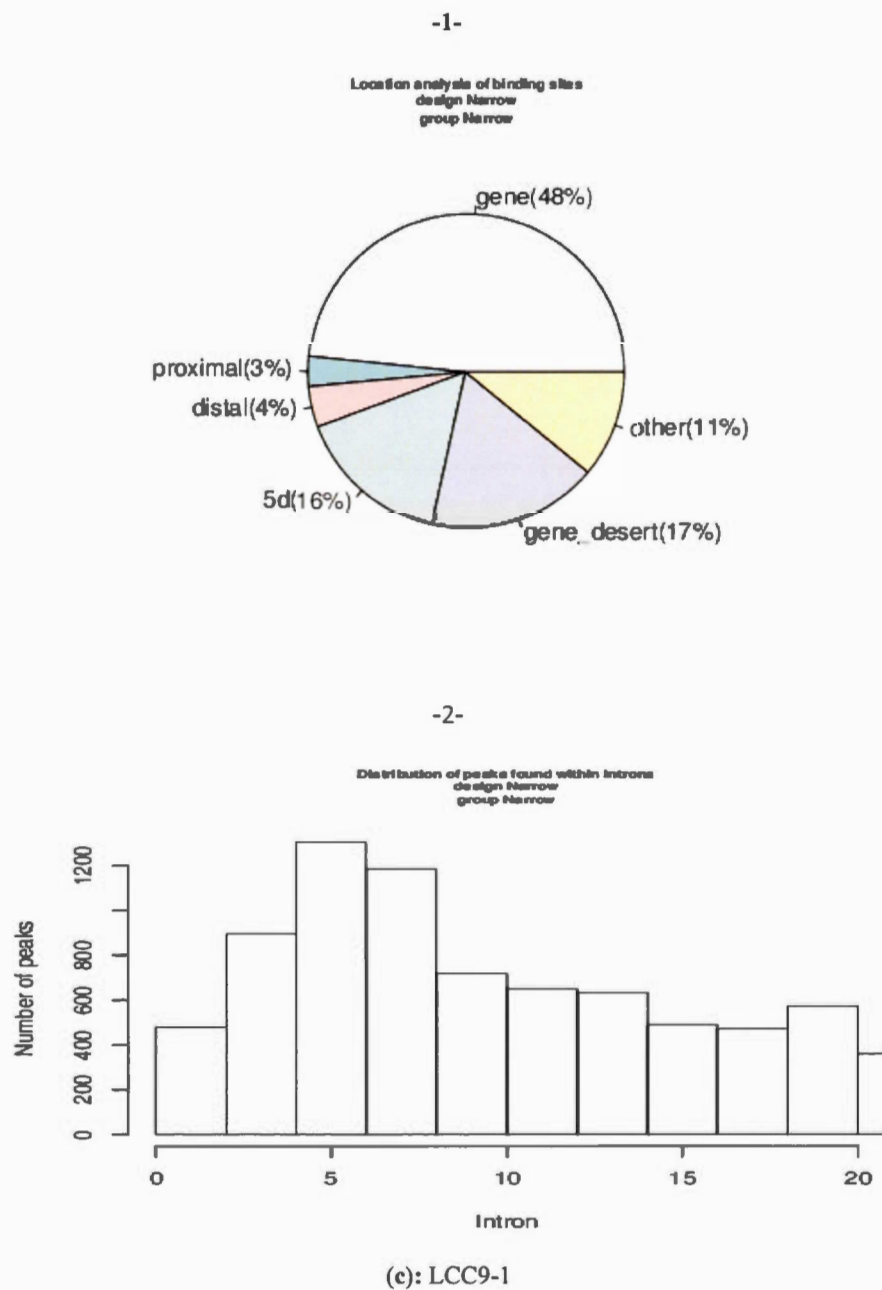
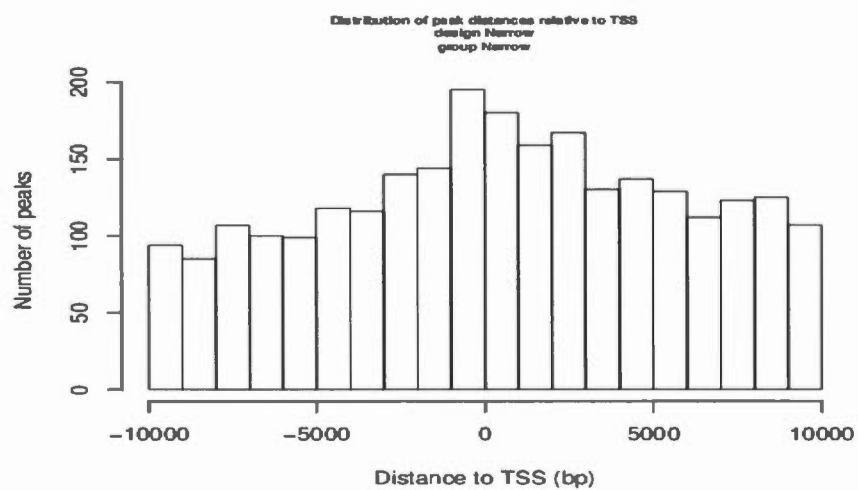


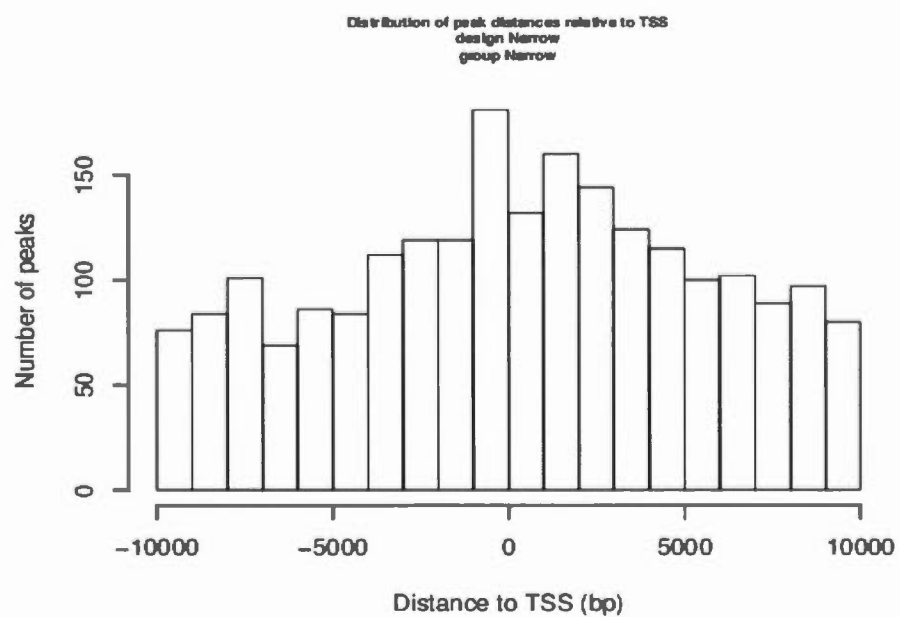
Figure 32: : Statistiques et annotations des localisations des pics par rapport aux gènes voisins les plus proches (site d'initiation de la transcription) dans chacune des conditions (a), (b) et (c): (a)1, (b)1

et (c)1 montrent les proportions des localisations génomiques des pics. Les localisations sont les suivantes: Gène (exon ou intron), proximal (0,2] kb en amont d'un site d'initiation de la transcription), Distal (2,10] kb en amont d'un site d'initiation de la transcription), 5d (10,100] kb en amont d'un site d'initiation de la transcription), Gene désert > = 100 kb en amont ou en aval d'un site d'initiation de la transcription) et Autre (non inclus dans les catégories ci-dessus). Seule une petite fraction, qui varie de 6 à 8 % selon l'expérience, se localise dans les régions promotrices proximal (0, 2] kb en amont du SST et distal (2, 10] kb en amont du SST. La plus grande fraction, variant de 48 à 51 % entre les expériences, réside dans les sites intra-géniques. (a)2 , (b)2 et (c)2 illustrent la distribution des pics trouvés dans les introns.

Sur la Figure 33, nous pouvons aussi observer la distribution du nombre des pics à proximité des sites de transcription (-10000, +10000 bp). Les figures montrent que les pics sont préférentiellement localisés dans les régions centrales où l'on trouve les distributions les plus élevées. Dans la lignée MCF7, les distributions du nombre de pics sont centrées et symétriques, de part et d'autre, par rapport au site d'initiation de la transcription (SST) et s'étalent de -5000 à + 5000 pb (voir Figure 33 (a)). Par contre, les distributions dans les lignées résistantes LCC2 et LCC9 affichent des profils similaires et sont plus importantes en amont (côté positif) du TSS (Figure 33(b) et (c), respectivement). Les pics se situent le plus souvent entre -1000 à + 5000 pb. Aussi, il est important de noter que ces profils de distribution concordent entre les deux replicats dans chaque lignée.

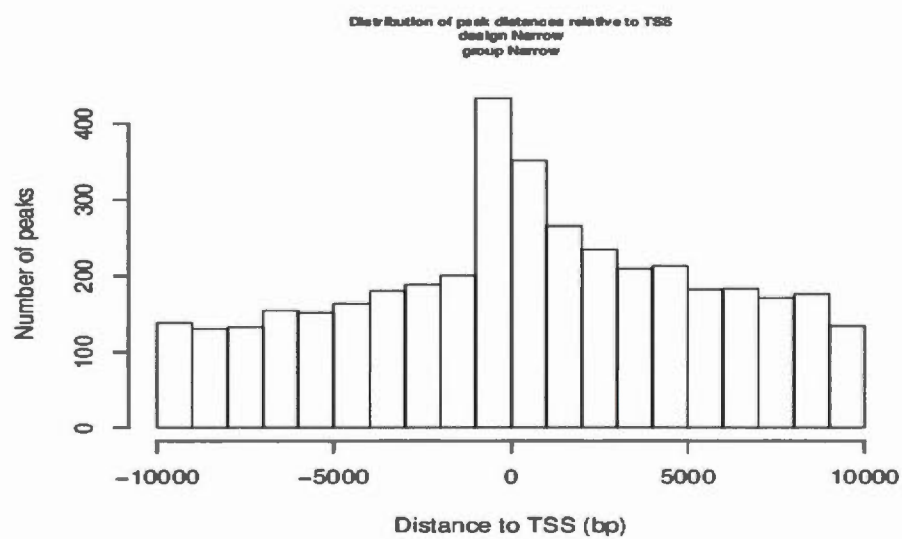


MCF7-1

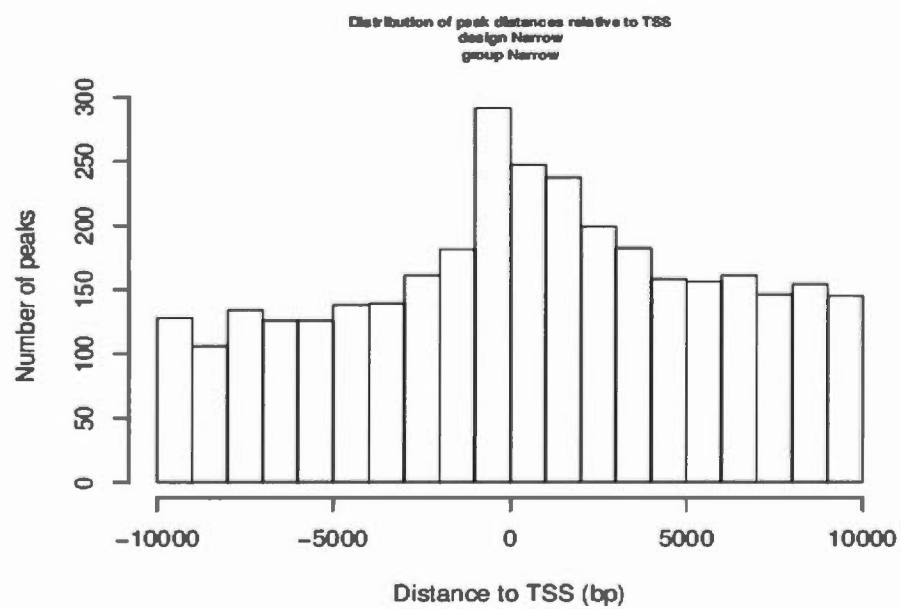


MCF7-2

(a)

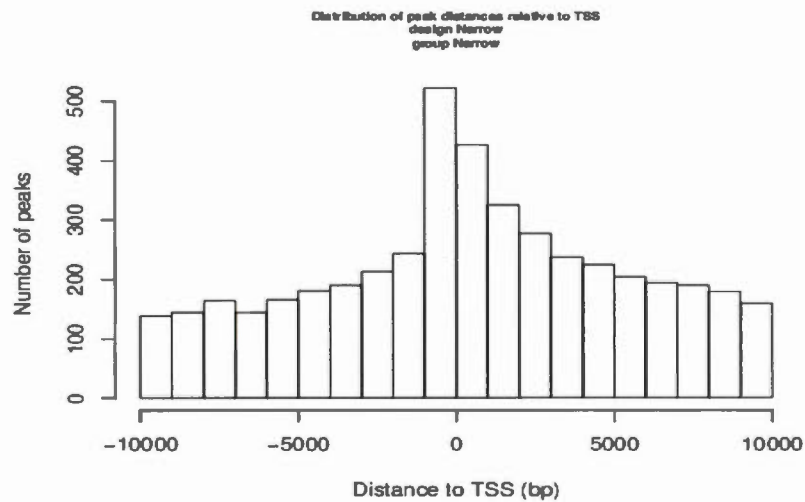


LCC2-1

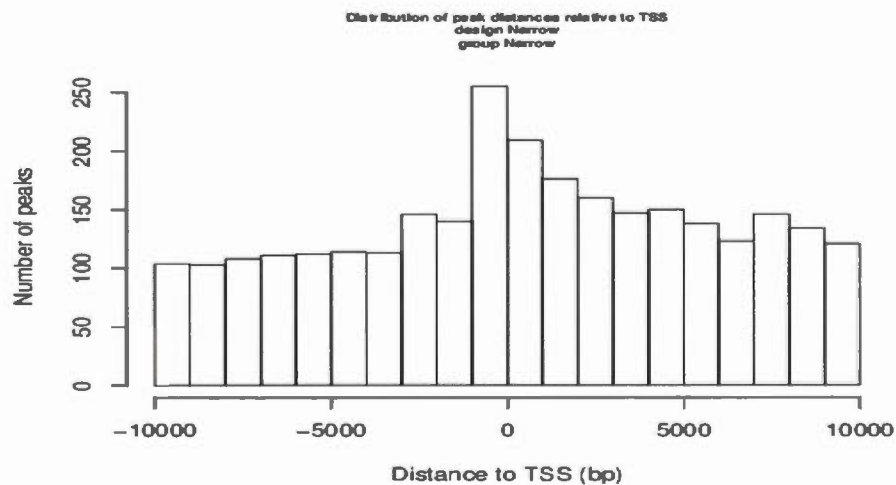


LCC2-2

(b)



LCC9-1



LCC9-2

(c)

Figure 33 : Distributions du nombre des pics à proximité des TSS (-10000, +10000 bp) pour chacune des expériences (incluant les expériences répliques): (a) (b) et (c) montrent que les pics sont préférentiellement localisés dans les régions centrales où l'on trouve les distributions les plus élevées. (a) dans la lignée MCF7, les distributions du nombre de pics sont centrées et symétriques de part et d'autre par rapport au site SST et s'étalent de -5000 à +5000 pb. (b) et (c) par contre, les distributions dans les lignées LCC2 et LCC9 affichent des profils similaires et sont plus importantes en amont (côté

positif) du SST. Les pics se situent le plus souvent entre -1000 à +5000 pb. (a), (b) et (c) les profils de distribution concordent entre les deux répliques dans chaque lignée.

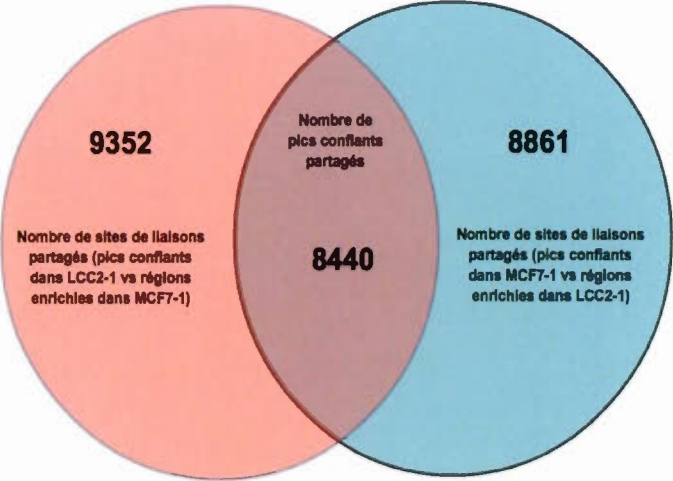
Identification des liaisons différentielles

- Taux de sites de liaison partagés et le taux des sites spécifiques à chaque condition

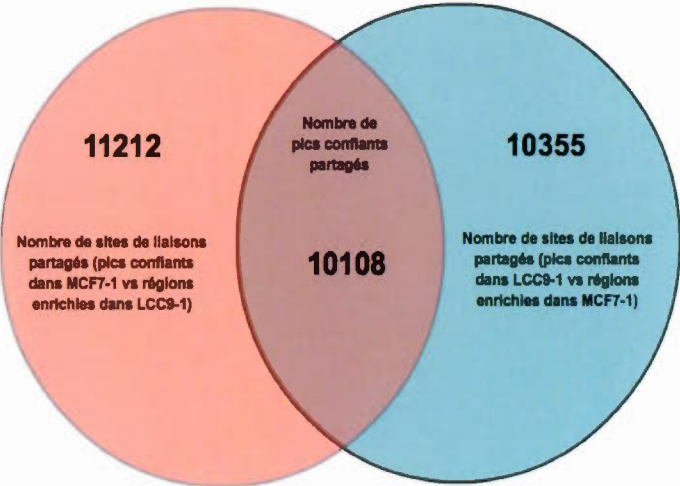
Notre méthode d'analyse des liaisons différentielles a été par la suite appliquée aux ensembles de pics des différentes expériences (conditions) issus du pipeline Génome Québec. Premièrement, nous avons comparé les profils des sites de liaison entre les différentes conditions deux à deux. Nos résultats (voir Tableau 13) ont montré 9773 sites de liaison partagés entre les cellules MCF7-1 et LCC2-1, 11459 sites entre MCF7-1 et LCC9-1 et 13915 sites entre LCC2-1 et LCC9-1.

Parmi ces sites partagés, 8440, 10108 et 12291 représentent des sites de liaison correspondant à des pics de haute qualité dans les deux conditions comparées, à savoir, MCF-7 vs LCC-2, MCF-7 vs LCC-9, et LCC-2 vs LCC-9, respectivement (voir Tableau 13 et Figure 34). Le reste des sites partagés présentent un pic de haute qualité (FDR 1%) dans l'une des conditions et une région significativement enrichie (mais qui se situe en dessous du seuil FDR de 1%) dans l'autre (voir les zones en dehors des intersections de la Figure 34). Afin de tester la cohérence entre répliques, les mêmes analyses ont été menées en utilisant une deuxième répliquât pour chaque condition. Malgré une légère différence dans les nombres de sites de liaison détectés entre les répliquât (voir Figure 31), une forte corrélation entre les résultats d'analyse, obtenus avec la première répliquât et la deuxième répliquât, a été observée (voir Figure S2 dans l'annexe).

En ce qui concerne les sites spécifiques aux conditions, l'analyse, contrairement à celle des sites partagés, est basée uniquement sur les sites des pics de haute qualité (FDR 1%). Autrement dit, même si une région enrichie de façon significative (mais située sous le seuil FDR 1%) est spécifique à une condition, elle ne sera pas considérée. Ceci est principalement dû au fait que dans les régions partagées, la confiance accordée au site identifié est assurée au moins par l'une des deux conditions sous comparaison. Par conséquent, le nombre de sites de liaison spécifiques à une condition donnée sera le nombre total de sites (pics) de haute confiance, auquel, on soustrait le nombre de sites de haute confiance qui présentent un enrichissement significatif (que se soient régions enrichies sous le seuil FDR ou pics de haute confiance) dans l'autre condition (voir Tableau 13). Ainsi, l'analyse a mis en évidence 5724 sites spécifique aux cellules MCF7-1 par rapport à LCC2-1, 9109 sites aux cellules LCC2-1 par rapport au cellules MCF7-1, 3373 sites aux cellules MCF7-1 par rapport aux cellules LCC9-1, 11898 sites aux cellules LCC9-1 par rapport aux cellules MCF7-1, 4816 sites spécifiques aux cellules LCC2-1 par rapport aux cellules LCC9-1 et 9692 sites spécifiques aux cellules LCC9-1 par rapport aux cellules LCC2-1. Ces résultats ainsi que ceux pour les expériences répliquas sont présentés dans le Tableau 13.



(a)



(b)

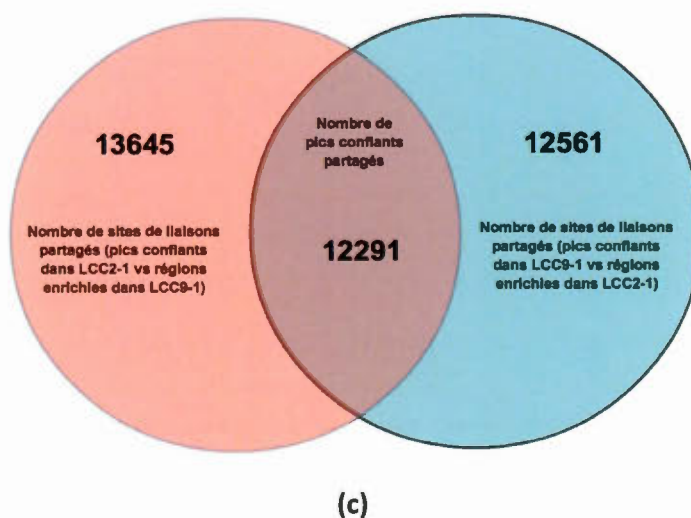


Figure 34 : Le nombre de site de liaisons partagées entre les trois conditions deux à deux (a), (b) et (c). Ces taux permettent d'avoir une idée globale des changements dans les régions liées, à savoir, le nombre de sites de liaison conservés, perdus ou acquis d'une condition à une autre.

Tableau 13 : Nombre de sites de liaison ERA identifiés dans les différentes conditions (FDR $\leq 1\%$)

	Condition	A vs B	B vs A	Nombre de sites de liaison partagés ((A vs B) U (B vs A))	Nombre de pics confiants partagés	Nombre total de pics confiants dans la condition	Nombre de sites de liaison spécifiques à la condition	Nombre de sites de liaison total dans la condition
A	MCF7-1	8861	9352	9773	8440	14585	5724	15497
B	LCC2-1					18461	9109	18882
A	MCF7-1	11212	10355	11459	10108	14585	3373	14832
B	LCC9-1					22253	11898	23357
A	LCC2-1	13645	12581	13915	12291	18461	4816	18731
B	LCC9-1					22253	9692	23807
A	MCF7-2	8052	8214	8771	7495	11783	3731	12502
B	LCC2-2					16295	8081	16852
A	MCF7-2	7296	8294	8665	6925	11783	4487	13152
B	LCC9-2					14090	5796	14481
A	LCC2-2	9606	10702	11162	9146	16295	6689	17851
B	LCC9-2					14090	3388	14550

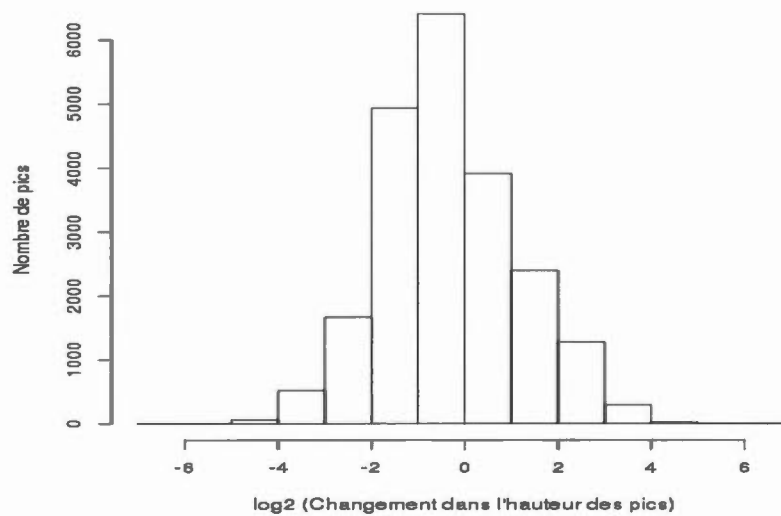
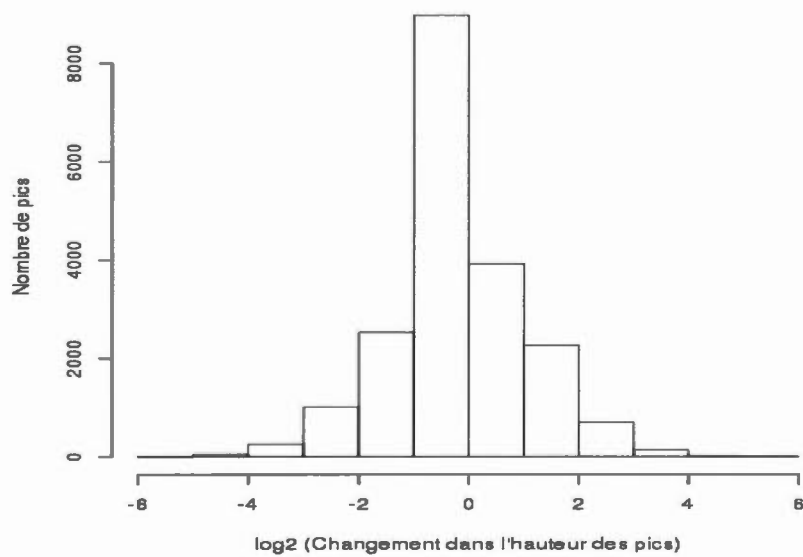
- nombre de sites de liaison spécifiques à la condition A=nombre total de pics de haute confiance (confiants) dans la condition A-(A vs B).
- nombre de sites de liaison total dans une condition A =nombre de sites de liaison partagés + nombre de sites de liaison spécifiques à la condition.

– Les liaisons différentielles (changements quantitatifs)

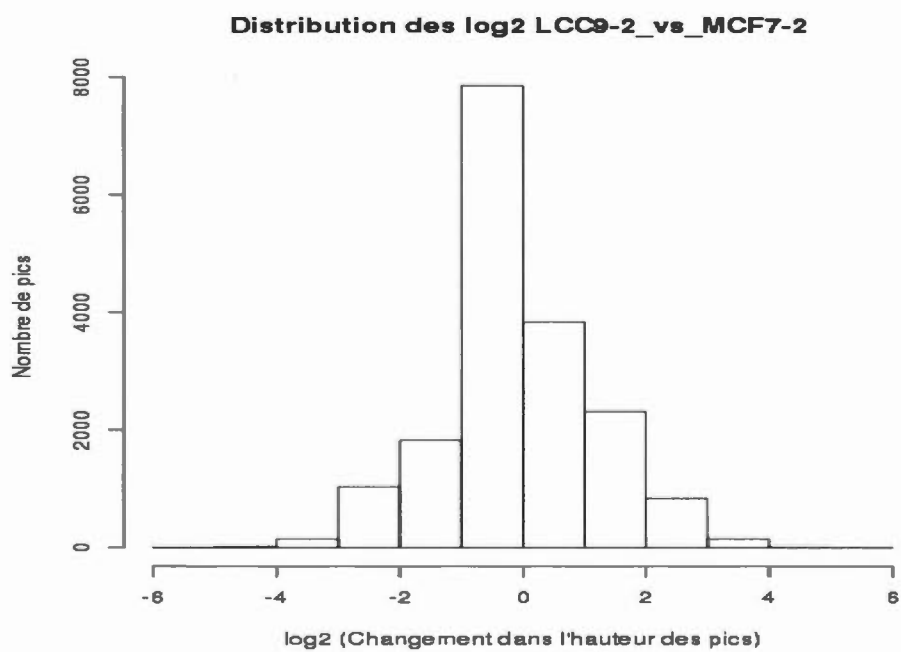
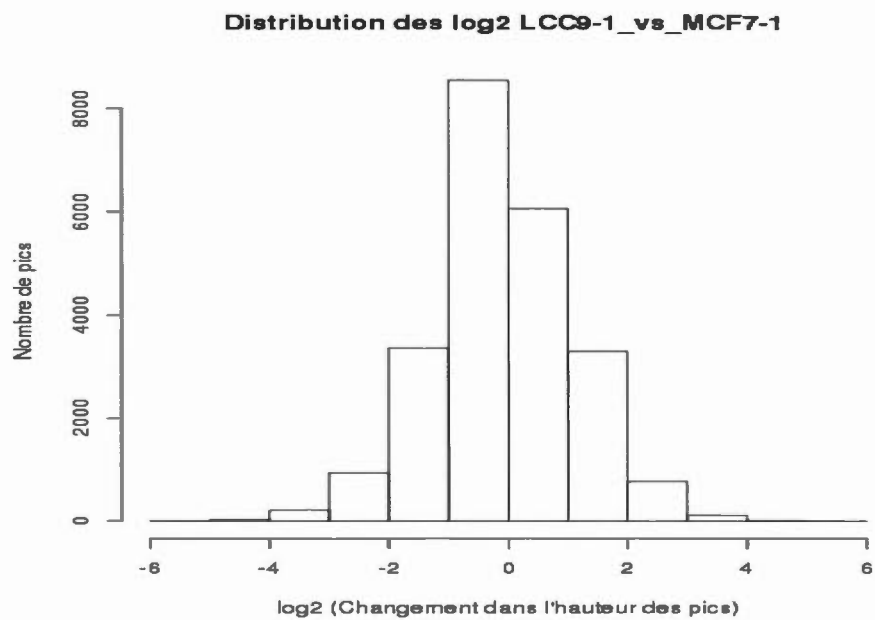
Bien que les sites de liaison spécifiques aux conditions, témoins de liaisons différentielles, soient bien mis en évidence dans la première étape, d'autres liaisons différentielles, résidant dans les sites de liaison partagés, ont besoin d'être identifiées. Ainsi, le but de l'étape d'analyse quantitative est de déterminer les changements

quantitatifs entre les hauteurs de pics au niveau des sites partagés. Ces derniers sont exprimés en termes de “fold change” log2, sur la base de la variation de la densité normalisée des lectures. Les résultats de l’analyse sont présentés dans la Figure 35.

Le premier élément enregistré sur les figures est que l’essentiel des pics (sites), et ce dans toutes les expériences de comparaison, sont dans l’intervalle ($-2 \text{ fold} < \text{score} < 2 \text{ fold}$) correspondant à la catégorie invariants. Les catégories croissants ($\text{score} > 2 \text{ fold}$) et décroissants ($\text{score} < -2 \text{ fold}$) sont, quant à elles, caractérisées par un petit nombre de pics (sites). Au total, l’analyse quantitative a révélé 1606 régions génomiques (sites) qui avaient significativement plus d’intensité de liaison de ERa dans les cellules résistantes LCC2 par rapport aux cellules MCF7 sensibles aux médicaments, et 2249 régions avec plus d’intensité de liaison ER dans les cellules sensibles MCF7 par rapport aux cellules LCC2 résistantes (Figure 35 (a)). De la même façon 893 régions génomiques (sites) avaient significativement plus d’intensité de liaison de ERa dans les cellules résistantes LCC9 par rapport aux cellules MCF7 sensibles aux médicaments, et 1180 régions avec plus d’intensité de liaison ER dans les cellules sensibles MCF7 par rapport aux cellules LCC9 résistantes (Figure 35 (b)). Enfin, 2973 régions génomiques (sites) avaient, et de façon significative, plus d’intensité de liaison de ERa dans les cellules résistantes LCC2 par rapport aux cellules LCC9 sensibles aux médicaments, et 2311 régions avec plus d’intensité de liaison ERa dans les cellules sensibles LCC9 par rapport aux cellules LCC2 résistantes (voir Figure 35 (c)).

Distribution des log2 LCC2-1_vs_MCF7-1**Distribution des log2 LCC2-2_vs_MCF7-2**

(a)



(b)

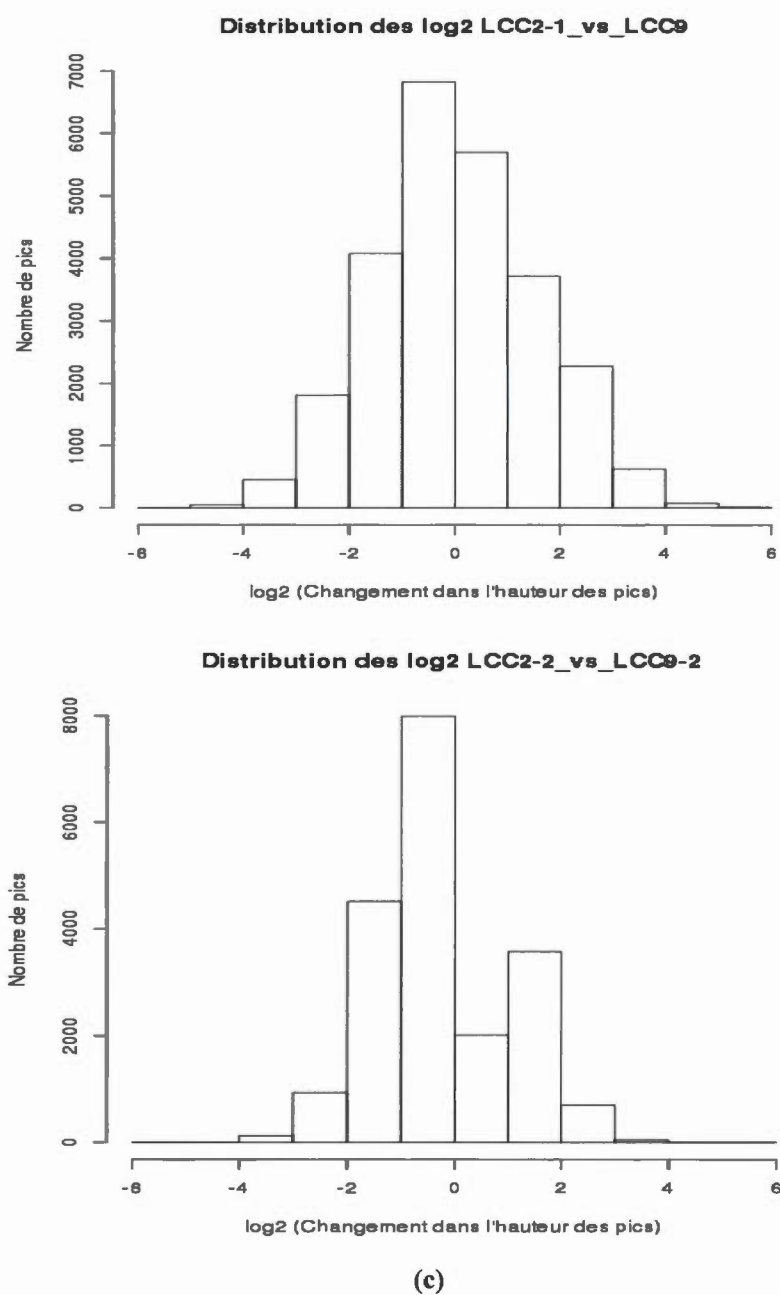


Figure 35 : Les résultats de l'analyse quantitative pour chaque paire de condition. Cette analyse détermine les changements quantitatifs entre les hauteurs de pics au niveau des sites partagés. Ceci est

exprimé en termes de "fold change" log2 sur la base de la densité normalisée des lectures. L'essentiel des pics dans toutes les expériences de comparaison (a) (b) et (c) sont dans l'intervalle (-2 fold < score < 2 fold) , qui correspond à la catégorie invariant. Le reste des pics est partagé entre la catégorie croissant (score > 2 fold) et la catégorie décroissant (score < -2 fold).

L'analyse de motifs

- Identification des motifs

À partir des pics identifiés par Macs et catégorisés dans l'étape de l'identification des liaisons différentielles, nous avons procédé à la recherche des motifs sur-représentés. Afin de vérifier si l'analyse par notre méthode a été correctement menée, nous avons, dans un premier temps, cherché à vérifier si les sites de liaison de notre facteur de transcription d'intérêt, ERa, ont été correctement identifiés dans chacun des jeux de données. Nous avons utilisé pour cela les fonctionnalités qu'offre HOMER, implanté dans le pipeline de génome québec mugqic.

Nous avons défini nos propres critères de sélection pour définir qu'un motif correspond à celui recherché. Pour cela, il faut que le site de liaison du facteur de transcription que nous recherchions apparaissent dans l'un des 3 meilleurs alignements et qu'il ait une valeur p-value inférieure à 10^{-5} . En appliquant ces deux filtres, nous avons identifié, dans chacun des jeux de données, la présence de motifs de séquence consensus dans les sites de liaison ERa représentant les motifs ERE d'intérêt (voir Figure 36 (a)), avec une très faible p-value et en tête de liste, comme voulu (voir exemple Figure 37).

De la même façon, la recherche de novo de motifs en utilisant le programme HOMER a également identifié des motifs raffinés ERE qui sont nettement similaires au ERE canonique (**GGTCAnnnTGACC**) (voir Figure 36 (b) et (c)). Un exemple de matrice de motif est aussi joint dans l'annexe).

AAGGTCA~~C~~CCG~~G~~TGACC

(a)

GGTCA~~C~~CCG~~G~~TGAC

(b)

GTCA~~G~~GGG~~T~~TGACC

(c)

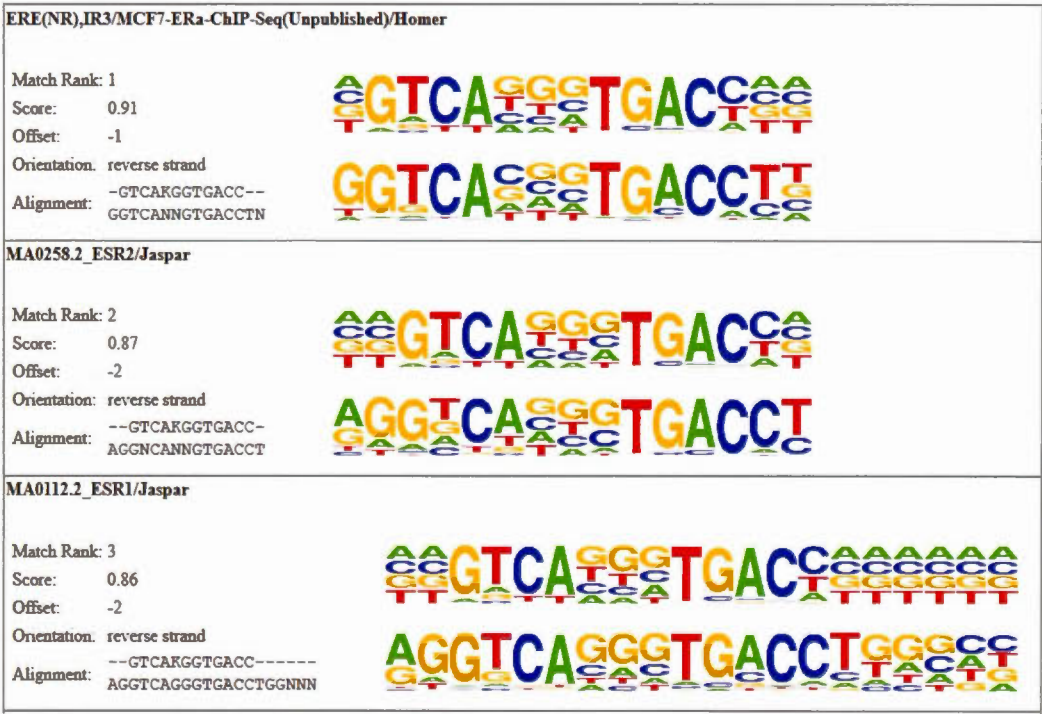
GTCA GGGTGACC

Reverse Opposite:

GGTCACCCGTGAC

p-value:	1e-55
log p-value:	-1.272e+02
Information Content per bp:	1.650
Number of Target Sequences with motif	126.0
Percentage of Target Sequences with motif	7.85%
Number of Background Sequences with motif	634.2
Percentage of Background Sequences with motif	1.31%
Average Position of motif in Targets	106.0 +/- 38.9bp
Average Position of motif in Background	101.0 +/- 63.4bp
Strand Bias (log2 ratio + to - strand density)	-0.2
Multiplicity (# of sites on avg that occur together)	1.07
Motif File:	file (matrix) reverse opposite
PDF Format Logos:	forward logo reverse opposite

(d)



(e)

Figure 36 : Les motifs consensus identifiés dans les sites de liaison ERE.

(a) ERE canonique trouvé dans la recherche de motifs connu. (b) et (c) deux exemples de motifs consensus, similaire au ERE canonique, identifié dans la recherche de novo. (d) Informations relatives au motif trouvé exemple du motif (c). (e) Résultats du top trois des alignements du motif (c) contre les motifs connus.

Total Target Sequences = 19247, Total Background Sequences = 30698













Rank	Motif	Name	P-value	log P-value	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Motif File	PDF
1		ERE(NR),IR3/MCF7-ERa-ChIP-Seq(Unpublished)/Homer	1e-1751	-4.033e+03	0.0000	3018.0	15.68%	558.7	1.82%	motif file (matrix)	pdf
2		FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-557	-1.283e+03	0.0000	6098.0	31.68%	5146.9	16.79%	motif file (matrix)	pdf
3		FOXA1(Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-552	-1.272e+03	0.0000	5394.0	28.02%	4304.0	14.04%	motif file (matrix)	pdf
4		Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1e-481	-1.109e+03	0.0000	4125.0	21.43%	3047.9	9.94%	motif file (matrix)	pdf
5		AP-2gamma(AP2)/MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer	1e-479	-1.104e+03	0.0000	3168.0	16.46%	2024.2	6.60%	motif file (matrix)	pdf
6		AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer	1e-452	-1.041e+03	0.0000	2616.0	13.59%	1535.4	5.01%	motif file (matrix)	pdf
7		Esrrb(NR)/mES-Esrrb-ChIP-Seq(GSE11431)/Homer	1e-392	-9.043e+02	0.0000	2668.0	13.86%	1716.2	5.60%	motif file (matrix)	pdf
8		FoxEbox(Forkhead,bHLH)/Panc1-Foxa2-ChIP-Seq(GSE47459)/Homer	1e-387	-8.918e+02	0.0000	4006.0	20.81%	3200.6	10.44%	motif file (matrix)	pdf
9		FOXP1(Forkhead)/H9-FOXP1-ChIP-Seq(GSE31006)/Homer	1e-196	-4.535e+02	0.0000	2063.0	10.72%	1611.5	5.26%	motif file (matrix)	pdf
10		GATA(Zf),IR3/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	1e-183	-4.229e+02	0.0000	1072.0	5.57%	617.8	2.02%	motif file (matrix)	pdf
11		GATA(Zf),IR4/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	1e-166	-3.824e+02	0.0000	618.0	3.21%	257.4	0.84%	motif file (matrix)	pdf
12		GATA3(Zf)/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	1e-161	-3.726e+02	0.0000	5417.0	28.14%	6120.2	19.97%	motif file (matrix)	pdf

Figure 37 : Extrait de la liste HTML des motifs enrichis (le top 12) obtenu avec le logiciel HOMER pour les pics dans la catégorie LCC2-1 vs MCF7-1 croissant ("increase").

- Identification des motifs secondaires

Après avoir identifié les motifs primaires attendus, correspondant à notre facteur de transcription, nous avons cherché si d'autres motifs biologiquement pertinents pouvaient être identifiés par HOMER. N'ayant pas de connaissances a priori concernant ces autres motifs, nous avons dû utiliser une approche différente. Pour cela, nous avons à nouveau fait appel à la fonction de visualisation des distributions des pourcentages de sites de liaison contenant le facteur de transcription collaborateur potentiel. Ainsi, en observant la distribution des pourcentages ainsi que le score des

alignements, nous sommes parvenus à identifier quatre motifs intéressants dans les jeux de données (voir Tableau 14)

Tableau 14 : Liste des motifs sur-représentés présents dans nos trois conditions avec le pourcentage des séquences cibles contenant ces motifs

		FOXP1	FOXA1	GATA	GATA-3	ERE	AP-2	AP-2 gamma	
LCC2-1_vs_LCC9-1	Decrease	16.1856	14.79	9.56333	10.94	9.96	10.95	12.38	
	Invariant	16.1856	14.79	9.56333	10.94	9.96	10.95	12.38	
	Increase	16.6789	15.4625	9.51333	10.81	15.75	16.19	17.73	
LCC2-2_vs_LCC9-2	Decrease	15.18	14.0325	9.22667	11.3967	11.71	15.345	16.02	
	Invariant	15.18	14.0325	9.22667	11.3967	11.71	15.345	16.02	
	Increase	16.1789	15.06	9.17333	10.5033	18.11	16.87	18.42	
LCC2-1_vs_MCF7-1	Decrease	15.1989	14.275	9.02222	10.2133	8.16	8.03	9.09	
	Invariant	15.1989	14.275	9.02222	10.2133	8.16	8.03	9.09	
	Increase	16.7156	15.5425	9.53556	10.85	15.68	15.025	16.46	
LCC2-2_vs_MCF7-2	Decrease	13.9422	12.4725	9.70222	10.9667	11.86	9.42	10.35	
	Invariant	13.9422	12.4725	9.70222	10.9667	11.86	9.42	10.35	
	Increase	15.7678	14.575	9.32111	10.5233	15.6	15.53	17.03	
LCC9-1_vs_MCF7-1	Decrease	15.5633	14.725	9.56889	10.3767	17.69	12.71	13.89	
	Invariant	15.5089	14.6775	9.61333	10.4267	17.66	12.77	13.95	
	Increase	16.4411	15.2675	9.36556	10.6867	15.37	15.2	16.65	
LCC9-2_vs_MCF7-2	Decrease	14.1567	12.8925	9.01778	9.86	14.3	11.245	12.43	
	Invariant	14.1978	12.9225	9.07444	9.89333	14.24	11.355	12.55	
	Increase	15.7367	14.51	9.22111	10.5067	16.99	15.695	17.19	

– Signification biologique des motifs enrichis trouvés

Nous avons par la suite fait une revue de la littérature afin de confirmer les interactions mises en évidence par notre méthode d'analyse. Ainsi, nous avons trouvé que les motifs Forkhead sont enrichis dans les régions liées par ERα, et plusieurs études ont identifié la protéine forkhead FoxA1 comme un facteur pionnier important pour l'interaction ER-chromatine (Laganiere, 2005; Carroll et al., 2005). De plus, FOXA1 est, lui-même, exprimé suite à un traitement par estrogène (Carroll, 2006). D'autres études ont aussi analysé la présence d'autres motifs enrichis dans les domaines de liaison de ER (Carroll et al., 2005) et ont révélé, en accord avec nos résultats, la surreprésentation des motifs pour les facteurs de transcription GATA (Lin et al., 2007). Similaire à FOXA1, GATA3 est un autre gène qui caractérise un cancer du sein ER + (Perou et al 2000 ; Sorlie et al., 2003). Son expression est corrélée à celle de ER (Lacroix et Leclercq, 2004) et des études récentes ont montré que ces deux protéines pouvaient être présentes au niveau des mêmes régions génomiques : 45% des sites de liaison de ER, dans les cellules cancéreuses mammaires MCF-7, recrutent également GATA3 (Kong *et al.*, 2011). Bien qu'une interaction physique entre GATA3 et ER n'ait pas été définie, l'hypothèse d'un rôle de GATA3 dans la signalisation ostrogénique a été confirmée. En effet, la déplétion de GATA3 entraîne une baisse de l'expression de ER (liée à un déficit de recrutement de l'ARN-PII sur le gène ESR1) et de l'expression des gènes estrogéno-dépendants comme SDF1 (Eeckhoute *et al.*, 2007). GATA3 est recruté au niveau des éléments de régulation cis de ER et est essentiel, entre autres, pour la transcription médiée par ER des gènes cibles, le maintien de la différenciation cellulaire chez l'adulte (Kouros-Mehr et al., 2006) et pour la croissance des cellules MCF-7 du cancer du sein (Eeckhoute et al., 2007). D'autres études ont aussi montré que FOXA1 et GATA3 sont en corrélation dans les cellules du cancer du sein (Yamaguchi et al, 2008) et ont longtemps été impliqués avec ER dans un même réseau (Lacroix et Leclercq 2004).

En ce qui concerne le motif AP-2, des résultats récents ont montré que ce dernier est surreprésenté dans les régions de liaison de ER (Tan et al. 2011). Il se lierait à une fraction significative de ces événements, où il serait essentiel pour la liaison de ER et la transcription médiée par l'estrogène (Tan et al. 2011). Fait intéressant, les régions co-liées par ER et AP-2 gamma ont aussi tendance à être occupées par FoxA1. Il semblerait aussi que AP-2 gamma et FoxA1 aient besoin l'un de l'autre pour une capacité de liaison efficace (Tan et al. 2011).

Ainsi, l'interaction entre ces protéines confirme que la découverte du site de liaison de FOXA1, GATA3 et AP-2 parmi les données de ER n'est pas le fruit du hasard.

Plus encore, la comparaison des pourcentages des séquences cibles, liées aux motifs, entre les conditions, montre une concordance avec les données retrouvées dans la littérature. Par exemple, nos résultats (voir tableau 14) montrent une association entre des niveaux élevés de AP-2 gamma et une réduction de la réponse à l'hormonothérapie. Une étude indépendante, qui a cherché spécifiquement une relation entre AP-2 gamma et la résistance à un traitement au tamoxifène, a montré des résultats similaires, en particulier chez les patientes post-ménopausées (Guler et al, 2007). L'expression de AP-2 gamma a été associée à une prolifération cellulaire continue, une progression de la maladie et une résistance à l'hormonothérapie. De la même façon, des études se sont penchées sur l'étude du facteur FOXA1 entre les cellules sensibles et les cellules résistantes à l'hormonothérapie. D' une part, les résultats ont montré que dans les cellules du cancer du sein sensibles, la dépendance au FOXA1 pour l'activité du tamoxifène sur ER est absolue. D'autre part, l'évaluation de la nécessité de FOXA1 dans la croissance des cellules de cancer du sein résistante au tamoxifène MCF-7(Tam-R), a permis de confirmer que malgré que la liaison de ER soit modifiée et indépendante du ligand, la croissance des cellules Tam-R nécessitent encore FOXA1 (Hurtado, 2011). Toutefois, les études concernant des changements dans les niveaux de FOXA1 entre les deux conditions sont

contradictoires. En effet, certaines avancent que le phénomène de résistance est accompagné de l'acquisition de régions de liaison au ER, et que ceci est corrélé avec un gain des liaisons au FoxA1 (Knowlden, 2003) (Hurtado, 2011) (Lupien, 2008). D'autres analyses révèlent un enrichissement statistique d'un nombre de séquences motifs incluant le motif FoxA1, mais seul un nombre mineur de co-occurrences des sites d'interaction ERa et FoxA1 a été détecté (Hurtado, 2011). Notre analyse de motifs enrichis a révélé que les motifs ERE montrent effectivement une liaison accrue dans les lignées cellulaires résistantes au tamoxifène (Tableau 14). Par contre, ils ne montrent pas de changements significatifs entre les conditions pour le facteur FOXA1.

Pour les motifs GATA enrichis dans les sites de liaison de ERa, nous n'avons noté que peu de variation entre les cellules sensibles et les cellules résistantes aux médicaments (Tableau 14). Des études (Ross-Innes, 2012) ont rapporté qu'ils montreraient un léger déclin lors de l'acquisition de la résistance aux médicaments, probablement en raison de la concurrence entre FOXA1 et GATA3.

Les annotations fonctionnelles des sites de liaison de ERa

Afin d'identifier les processus biologiques et les voies modifiées par ERa et dans le but d'effectuer l'analyse d'enrichissement pour les fonctions des gènes (les gènes voisins les plus proches des sites de liaison de ERa), nous avons utilisé le système d'annotation des fonctions des catégories de Gene Ontology (GO). Les termes GO enrichis et statistiquement significatifs (test Hypergéométrique, P-value < 0.01) ont été identifiés en utilisant l'outil HOMER, intégré dans le pipeline génome québec.

De façon générale, les termes enrichis, liés aux gènes assignés aux sites des liaisons invariables entre les cellules MCF7 sensibles et les cellules résistantes LCC-2, et les termes enrichis, liés aux gènes assignés aux sites des liaisons différentielles avec une importante intensité du signal dans les cellules MCF7, sont identiques. Par contre, les

termes enrichis, liés aux gènes assignés aux sites caractérisés par des liaisons différentielles avec augmentation de l'intensité de signal dans les cellules LCC2, sont totalement différents ou présentent quelques entités en commun avec les autres sites. La même observation a été faite entre les cellules LCC9 et MCF-7 et entre les cellules LCC2, LCC9 (voir tableau 2S en annexe). Toutefois, bien qu'ils ne soient pas identiques, ces termes appartiennent majoritairement aux mêmes familles. Dans ce qui suit, une description plus détaillée, pour les comparaisons deux à deux, est présentée.

LCC-2 vs MCF-7

Les gènes assignés aux sites, présentant des liaisons invariables entre MCF-7 et LCC-2, et les sites de liaisons différentielles, avec intensité de signal dans les cellules MCF-7, sont principalement associés à la **régulation biologique, le développement, les processus métaboliques et les processus cellulaires** dans les termes des processus biologiques, **la liaison des protéines** dans les termes des fonctions moléculaires, **la membrane** ainsi que **la jonction cellulaire** dans les termes des composants cellulaires. En ce qui concerne les gènes assignés aux sites de liaison avec une intensité du signal dans les cellules LCC-2 par rapport à MCF-7 (liaisons différentielles), ils seraient associés à **la régulation biologique, les processus du développement et la localisation** dans les termes des processus biologiques, **la liaison de protéine** (mais pas les mêmes que dans les autres sites partagés avec les cellules MCF-7) dans les termes des fonctions moléculaires, et **la membrane, la jonction cellulaire et la projection cellulaire** dans les termes des composants cellulaires.

LCC-9 vs MCF-7

Les gènes assignés aux sites, présentant des liaisons invariables entre LCC-9 et MCF-7, et les sites de liaisons différentielles, avec intensité de signal dans les cellules MCF-7, sont principalement associés au **développement** dans les termes des processus biologiques, **l'activité et la liaison des protéines** dans les termes des fonctions moléculaires, et la **membrane ainsi que la jonction cellulaire** dans les termes des composants cellulaires. Pour les gènes assignés aux sites de liaison avec une intensité du signal dans les cellules LCC-9 par rapport à MCF-7 (liaisons différentielles), ils seraient associés avec la **régulation biologique, processus du développement et la localisation** dans les termes des processus biologiques, la **liaison de protéine** (mais pas les mêmes que dans les autres sites partagés avec les cellules MCF-7) dans les termes des fonctions moléculaires, et la **membrane, la jonction cellulaire et la vésicule** dans les termes des composants cellulaires.

LCC-2 vs LCC-9

Les gènes assignés aux sites, présentant des liaisons invariables entre LCC-2 et LCC-9, et les sites de liaisons différentielles, avec intensité de signal dans les cellules MCF-7, sont principalement associés au **développement, la localisation, la migration et la motilité cellulaire** dans les termes des processus biologiques, **l'activité et la liaison des protéines et l'activité du potassium** dans les termes des fonctions moléculaires, et la **membrane** dans les termes des composants cellulaires. Pour les gènes assignés aux sites de liaisons, avec une intensité du signal dans les cellules LCC-9 par rapport à MCF-7 (liaisons différentielles), ils seraient associés à la **régulation biologique et les processus du développement** dans les termes des processus biologiques, la **liaison et l'activité de protéine** (mais pas les mêmes que dans les autres sites partagés avec les cellules MCF-7) dans les termes des fonctions

moléculaires, et **la membrane, la jonction cellulaire, la projection cellulaire et la vésicule** dans les termes des composants cellulaires.

Les fonctions de gène, pour tous les gènes voisins les plus proches, sont résumées dans le tableau S2 (on a pris les 10 premiers termes enrichis, voir S2 de l'annexe).

Nous avons aussi observé un enrichissement des termes COSMIC (Catalogue Of Somatic Mutations In Cancer), qui donnent une vue sur les gènes mutés dans des cancers similaires, mais aussi un enrichissement des termes des listes de la base de données des signatures moléculaires MSigDB ("**Genes sets for pathways, factor/miRNA target predictions, expression patterns, etc.**"). Ainsi, dans l'ensemble des lignées, au niveau de tous les sites, qu'ils soient des sites présentant des liaisons différentielles (et indépendamment du type cellulaire contenant le signal) ou des sites avec liaisons invariables, les gènes sont essentiellement associés aux termes des tumeurs endocrines dans les termes COSMIC et à des signatures moléculaires liées aux termes de cancer de sein et de la résistance aux tamoxifène (voir les termes détaillés dans le tableau 2S dans l'annexe).

La base de données des voies KEGG a aussi été utilisée pour identifier des modules de fonctions régulées par ERa. Le top 10 des voies significativement enrichies (P -value < 0.01) est présenté (Tableau 3S en annexe) :

LCC-9 vs MCF-7

Les voies dans le cancer, la régulation de cytosquelette d'actine, la voie de signalisation Rap1, la voie de signalisation Ras, la voie de signalisation PI3K-Akt, la voie de signalisation cGMP-PKG et la voie Hippo sont classées parmi les voies les plus enrichies liées aux gènes assignés aux sites présentant des liaisons invariables entre les deux types cellulaires ou des liaisons différentielles avec une intensité du

signal dans les cellules MCF-7. Concernant les sites qui abritent des liaisons différentielles avec intensité du signal dans les cellules LCC-9, le top des voies enrichies inclue la voie d'adhésion focale, la voie de guidance axonal, la voie de signalisation Rap1, les voies du cancer, et la voie de signalisation ErbB.

LCC-2 vs MCF-7

Les voies dans le cancer, les MicroRNAs dans le cancer, la voie de signalisation HIF-1 sont classées parmi les voies les plus enrichies liées aux gènes assignés aux sites présentant des liaisons invariables entre les deux types cellulaires ou des liaisons différentielles avec une intensité du signal dans les cellules MCF-7. Concernant les sites qui abritent des liaisons différentielles avec intensité du signal dans les cellules LCC-2, le top des voies enrichies est la voie d'adhésion focale, la voie de guidance axonal, la voie de signalisation Rap1, les voies du cancer, et la voie de signalisation cGMP-PKG.

LCC-2 vs LCC-9

Les voies de guidance axonal, les voies de signalisation de Chemokine, les voies de signalisation Hippo, le métabolisme de Inositol phosphate, et les protéoglycanes dans le cancer sont classées parmi les voies les plus enrichies liées aux gènes assignés aux sites présentant des liaisons invariables entre les deux types cellulaires. Au niveau des sites présentant des liaisons différentielles avec une intensité du signal dans les cellules LCC-9, les voies les plus enrichies sont celles de guidance axonal, les voies de signalisation Hippo, les MicroRNAs dans le cancer, les voies de signalisation HIF-1. Concernant les sites qui abritent des liaisons différentielles avec intensité du signal dans les cellules LCC-2, le top des voies enrichies inclue la voie d'adhésion focale, la voie de guidance axonal, la voie de signalisation Rap1, les voies du cancer, les voies de signalisation Hippo et la voie de signalisation ErbB.

– La signification biologique des voies enrichies

Nous avons effectué une revue de la littérature afin d'évaluer le degré de signification des résultats obtenus des annotations. Le constat est que le top des maps enrichis, comme la voie d'adhésion focale, la voie de signalisation Rap1, la voie de signalisation Ras, la voie de signalisation HIF-1, la voie de signalisation PI3K-Akt, la voie de signalisation cGMP-PKG et la voie de signalisation ErbB, ont été tous rapportés comme étant liés à ERa dans le cancer du sein. Plus de détails sur chacune de ces voies sont présentés dans les sections suivantes :

• La kinase d'adhésion focale

Une expression élevée de la kinase d'adhésion focale a été rapportée comme étant liée à la progression du cancer du sein, et les tumeurs avec une expression élevée de cette kinase perdent ER et PR (Lark, 2005.).

• La voie de signalisation Rap1

L'activation de Rap1 augmente la migration et l'invasion des cellules du cancer de la prostate et du cancer du sein. Par contre, l'inhibition de l'activité Rap1 par le knock-down médié par ARNi ou l'expression ectopique de Rap1GAP affecte la migration et l'invasion des cellules cancéreuses. Dans un modèle de xénogreffe de souris, le blocage de la signalisation Rap1 par expression de Rap1GAP a détruit les métastases du cancer du sein. Ces études soutiennent un rôle pour l'activation altérée de Rap1 dans la progression des métastases du cancer du sein et de la prostate, et suggèrent

que le ciblage de la signalisation Rap1 pourrait fournir un moyen de contrôler les métastases de ces cancers (Itoh, 2007).

- La voie de signalisation Ras

Bien que des formes mutées de Ras ne soient pas associées à la majorité des cancers du sein (< 5%), il existe de considérables preuves expérimentales que l'hyperactivation de Ras peut favoriser la croissance et le développement du cancer du sein. Un travail (Eckert, 2004) a constaté que cinq des neuf lignées cellulaires du cancer du sein utilisées dans l'étude montrent des niveaux élevés de Ras-GTP actifs qui peuvent être, en partie, dus à l'activation de HER2.

- La voie de signalisation cGMP-PKG

Des données (Schwappacher, 2013) suggèrent que PKGI β amplifie la motilité de cellules de cancer du sein et de la capacité invasive, au moins en partie, par la phosphorylation d'une protéine appelée CaD. Ces résultats ont ainsi permis d'identifier une fonction pro-migratoire et pro-invasive pour PKGI β dans les cellules du cancer du sein humain, ce qui suggère que PKGI β est une cible potentielle pour le traitement du cancer du sein.

- La voie de signalisation PI3K-Akt

Akt est une sérine / thréonine kinase dont l'étude a démontrée qu'elle jouait un rôle important dans la survie, lorsque les cellules sont exposées à différents stimuli apoptotiques. Des études récentes ont montré que l'activation altérée d'Akt dans le carcinome du sein est associée à un mauvais pronostic et à la résistance à l'hormonothérapie et la chimiothérapie (Ciruelos, 2014). La voie de signalisation Akt

attire actuellement une attention considérable en tant que nouvelle cible pour des stratégies thérapeutiques efficaces. Les résultats d'une étude (Tokunaga, 2006) suggèrent que l'activation d'Akt induirait une résistance endocrine dans le cancer du sein métastatique, quel que soit le type d'agent endocrine qui a été administré. Les résultats indiquent, aussi, que l'activation d'Akt en aval de la voie de HER2 joue un rôle important dans la résistance à l'hormonothérapie pour le cancer du sein. Cela suggère que pAkt peut être un prédicteur utile de la résistance à la thérapie endocrine du cancer du sein, et que son inhibition pourrait augmenter l'efficacité de la thérapie.

- La voie de signalisation ErbB

ErbB2 est considéré comme un récepteur tyrosine kinase (RTK) transmembranaire classique présentant une activité de tyrosine-kinase sans ligand, de haute affinité, connu. L'amplification du gène *ErbB2* et / ou la surexpression de l'ARNm ou de la protéine se produit dans environ 20% des cancers du sein. C'est un témoin d'une maladie agressive avec une évolution clinique de mauvais diagnostic (Arteaga, 2012). Les complexes ErbB homo- et hétérodimériques initient un spectre varié de signalisation, engageant des voies telles que la voie PI3K / Akt et les voies Ras / Raf / MEK / MAPK (Yarden, 2001) (la voie PI3K / Akt et la voie Ras, étant rapportées par notre analyse). En outre, la spécificité et la puissance des sorties de signalisation sont déterminées par la multitude d'adaptateurs et d'enzymes qui s'associent avec les complexes de récepteurs.

- La voie de signalisation Hippo

La voie Hippo, qui contrôle la taille des organes dans les diverses espèces, est aussi démontrée comme un réseau de signalisation du cancer. La dérégulation de cette voie

peut induire des tumeurs dans des organismes modèles, et se produit dans un large éventail de cancers humains, notamment du poumon, du côlon, des ovaires et le cancer du foie (Harvey, 2013).

- Le guidage axonal

Le guidage axonal a joué des rôles importants dans les cancers. Des molécules de guidage axonal peuvent contrôler le développement, la migration et l'invasion des cellules cancéreuses (Chedotal, 2007).

- La régulation du cytosquelette d'actine

La régulation du cytosquelette d'actine est liée à la migration et l'invasion des cellules du cancer (Yamaguchi, 2007).

- La voie de signalisation HIF-1 α

Des niveaux élevés de HIF-1 α sont associés à une augmentation de la prolifération et de l'expression de ER et le VEGF. Ainsi, des niveaux accrus de HIF-1 α sont potentiellement associés à des tumeurs plus agressives (Bos, 2000).

- La voie de métabolisme du phosphate d'inositol

Les membres de la voie du métabolisme du phosphate d'inositol régulent la prolifération cellulaire, la migration et la signalisation phosphatidylinositol-3-kinase (PI3K) / Akt, et sont fréquemment dérégulés dans le cancer (Tan, 2015).

- La voie de signalisation des chimiokines

Les chimiokines sont des facteurs solubles qui ont été démontré jouant un rôle important dans la régulation du recrutement des cellules immunitaires au cours des réponses inflammatoires et de la défense contre les agents pathogènes étrangers. L'expression ou l'activité dérégulée de plusieurs voies de signalisation de chimiokines ont été impliquées dans la progression du cancer. Alors que les études dans le passé ont mis l'accent sur le rôle de ces voies de signalisation de la chimiokine dans la régulation des réponses immunitaires, des études montrent que ces nouvelles molécules régulent divers processus cellulaires, y compris l'angiogenèse, et la régulation de la croissance des cellules épithéliales et la survie. De nouvelles preuves indiquent que les chimiokines sont indispensables pour la progression du cancer et montrent des fonctions complexes et diversifiées dans le microenvironnement de la tumeur (Hembruff, 2009).

- Les protéoglycans

Les protéoglycans contrôlent de nombreux processus normaux et pathologiques, parmi lesquels la morphogenèse, la réparation des tissus, l'inflammation, la vascularisation et la métastase du cancer. Au cours du développement de la tumeur et la croissance, l'expression des protéoglycans est nettement modifiée dans le microenvironnement tumoral. L'expression altérée des protéoglycans au niveau des membranes cellulaires tumorales et stromales affecte la signalisation des cellules du cancer, la survie et la croissance, l'adhésion cellulaire, la migration et l'angiogenèse. Malgré la grande complexité et l'hétérogénéité du cancer du sein, l'évolution rapide de notre connaissance est que les protéoglycans sont parmi les principaux acteurs dans le microenvironnement de la tumeur du sein, ce qui suggère leur potentiel comme cibles pharmacologiques dans ce type de cancer. Il a été récemment suggéré

que le traitement pharmacologique peut cibler le métabolisme des protéoglycanes, leur utilisation comme cibles pour l'immunothérapie ou leur utilisation directe comme agents thérapeutiques (Theocharis, 2015).

Toutes ces annotations confirment le rôle crucial des ERs dans le développement, la migration et l'invasion du cancer du sein.

CONCLUSION

Ce mémoire présente deux contributions importantes. En premier, le développement et l'implémentation d'une plateforme bio-informatique pour l'analyse différentielle des données ChIP-Seq liées à un facteur de transcription entre plusieurs conditions (plusieurs lignées cellulaires, plusieurs étapes de développement, etc.). Cette plateforme implantée sur MUGQIC sera disponible et pourra faire partie intégrante du projet GenAP. En second, il présente des résultats originaux sur le facteur de transcription ERa, qui est un marqueur moléculaire important du cancer du sein. Les données proviennent de trois lignées cellulaires : la lignée MCF7 sensible aux médicaments (le tamoxifène et le fulvestrant), la lignée LCC2 résistante au tamoxifène et la lignée LCC9 résistante au tamoxifène et au fulvestrant. Afin de récapituler et tester la concordance des résultats observés tout au long de l'analyse, l'expérience ChIP-seq de ERa a été réalisée en deux exemplaires (réplicats), pour chacune des trois lignées cellulaires.

Les fragments de départ, issus des expériences ChIP-Seq, étaient déjà alignés sur le génome humain hg18. Un prétraitement a été effectué pour les convertir et les réaligner sur la version la plus récente hg19. Les régions du génome présentant un enrichissement significatif en séquences alignées (pics) ont été par la suite identifiées, dans chacune des bibliothèques, en appliquant un logiciel d'appel de pic "peak caller" ($\text{FDR} \leq 1\%$). À cette étape, nous avons identifié un total de 14 586, 18 462 et 22 254 sites de liaison de ERa de haute confiance ($\text{FDR} \leq 1\%$) dans les cellules MCF7-1, LCC2-1 et LCC9-1, respectivement. Dans les expériences répliques MCF7-2, LCC2-2 et LCC9-2, un total de 11 784, 16 296 et 14 091 sites de liaison ont été identifiés, respectivement. L'analyse de la localisation génomique des sites de liaison, par

rapport aux gènes voisins les plus proches, a révélé que seule une petite fraction des pics, de 6 à 8% selon l'expérience, se localisaient dans les régions promotrices. Cela suggère que le facteur de transcription ERa est impliqué dans le contrôle à longue portée. Des travaux antérieurs ont montré une abondance des sites de liaison ERa localisés loin des promoteurs (Lin et al, 2007) et avaient validé que ERa comprenait le "looping" (Carroll et al, 2006). Nos résultats ont fourni un soutien supplémentaire pour ce mode de fonctionnement du facteur ERa dans les trois lignées cellulaires.

Dans l'étape suivante, les régions enrichies ont été soumises à l'analyse comparative afin de caractériser les liaisons différentielles dans chaque paire de condition. Cette dernière comporte deux types d'analyse : une analyse qualitative et une autre comptitative. L'analyse qualitative a donné une évaluation globale du nombre des sites de liaison partagés et non partagés entre les conditions deux à deux. Une stratégie de combinaison des seuils ($FDR \leq 1\%$) et ($P\text{-value} < 10^{-5}$) pour les sites de liaison a été appliquée. Cette approche originale a permis de filtrer les faux positifs dans l'ensemble des sites spécifiques aux conditions. Ainsi, à ce niveau, un ensemble noyau de 9773, 11459 et 13915 sites partagés entre la lignée LCC2 et MCF7, LCC9 et MCF7 et LCC2 et LCC9 ont été identifiés, respectivement. Concernant les sites spécifiques à chaque condition, 5724 et 9109 sites de liaison spécifiques à MCF7 et LCC2, respectivement, ont été identifiés dans la comparaison LCC2 vs MCF7, 11898 et 3373 sites spécifique à LCC9 et MCF7, respectivement, dans la comparaison LCC9 vs MCF7 et un total de 4816 et 9692 sites de liaison spécifiques à LCC2 et LCC9, respectivement, dans la comparaison LCC2 vs LCC9.

L'analyse quantitative, quant à elle, a fourni des mesures pour les changements quantitatifs entre les conditions. La densité des reads enregistrée au niveau de chaque site de liaison a été mesurée et comparée entre les conditions afin d'en extraire les liaisons différentielles. Dans la paire LCC2 vs MCF7, 1606 sites de liaison avaient significativement plus d'intensité de liaison de ERa dans les cellules LCC2 par rapport aux cellules MCF7, et 2249 sites avec plus d'intensité de liaison ERa dans les

cellules MCF7 par rapport aux cellules LCC2. Dans la paire LCC9 vs MCF7, 893 sites avec significativement plus d'intensité de liaison de ERa dans les cellules LCC9 par rapport aux cellules MCF7, et 1180 sites avec plus d'intensité de liaison ERa dans les cellules MCF7 par rapport aux cellules LCC9. Dans la paire LCC2 vs LCC9, 2973 sites avaient significativement plus d'intensité de liaison de ERa dans les cellules LCC2 par rapport aux cellules LCC9, et 2311 sites avec plus d'intensité de liaison ERa dans les cellules LCC9 par rapport aux cellules LCC2. Enfin, une analyse de motifs et des annotations fonctionnelles des sites de liaisons, caractérisés et classés dans des ensembles, ont aussi été faites.

En résumé, tous ces résultats montrent que le récepteur ERa se lie au niveau de toutes les lignées cellulaires (ou conditions), mais c'est le nombre des événements de liaison identifiés, leur localisation et l'intensité des pics au niveau des sites qui diffèrent entre les cellules sensibles et les cellules résistantes aux médicaments. Les mêmes résultats concernent aussi la comparaison entre les deux lignées résistantes aux médicaments. Ces observations confirment que la résistance aux médicaments n'est pas due à la perte de la liaison ERa à l'ADN (Lupien, 2008). De plus, nous avons non seulement retrouvé, dans chacun des jeux de données sous comparaison, le site de liaison du facteur de transcription d'intérêt ERa, mais nous avons également réussi à identifier des motifs rapportés dans la littérature comme se liant à la même région de liaison que ERa. Ces motifs apparaissent, selon la littérature, à des niveaux variables selon la condition (résistance ou pas aux médicaments), en accord avec nos résultats. L'analyse de l'enrichissement de GO a donné un aperçu de la fonction des gènes. Ces résultats ont indiqué que les gènes régulés par ERa sont liés au développement, la progression et les métastases du cancer du sein. En outre, l'analyse des voies KEGG ont aussi montré que les gènes régulés par ERa sont enrichis dans certaines voies liées aux maladies (les voies dans le cancer étant la voie KEGG la plus enrichie),

d'autres au système immunitaire, comme la voie de signalisation de la chimiokine, ainsi qu'aux voies de signalisation en lien avec les facteurs de croissance HER2.

Les différences dans les liaisons ER entre contextes sensibles et résistantes peuvent être dues à une reprogrammation de liaison ER suite à des stimuli spécifiques. Les voies de facteurs de croissance ont longtemps été impliquées dans la modulation de ces stimuli mais plus d'études sont nécessaires pour mieux comprendre les mécanismes.

Pour répondre aux défis posés par l'analyse d'un ensemble d'expériences ChIP-Seq, plusieurs solutions ont été apportées dans cette étude. Du côté technique, pendant notre étude, le pipeline MUGQIC était sous construction continue et contenait de nombreux bugs que nous avons dû régler avant de permettre sa configuration et son utilisation. D'autre part, en offrant l'avantage d'une double analyse comparative dans un même environnement, notre méthode est une solution pour mener des analyses complètes. Par ailleurs, nous avons ajouté une étape de prétraitement de données par rapport au pipeline standard décrit dans la littérature. Cette étape est en mesure de convertir des données, déjà alignées sur des versions anciennes du génome d'intérêt, de les traiter puis de les réaligner sur la version la plus récente du même génome.

Enfin, de façon générale, notre plateforme d'analyse intégrée, accessible à partir de Calcul Québec, se distingue par son approche originale et efficace pour traiter des données ChIP-Seq et mener des analyses comparatives entre plusieurs conditions. Sa partie analyse qualitative offre une meilleure précision pour l'identification et la distinction entre les sites de liaison partagés et spécifiques aux conditions, tandis que l'analyse quantitative fournit le détail de l'intensité du signal (hauteur de pics) de liaison au niveau de ces sites. De plus, les sites sont catégorisés de manière à faciliter l'interprétation des résultats, au sein de chaque catégorie et entre catégories, lors des analyses en aval. Enfin, avec notre approche, plusieurs nouveaux sites de liaisons pour le facteur ERα ont été mis en évidence. Ces résultats seront étudiés, confirmés et expérimentés en laboratoire prochainement.

GLOSSAIRE

ADN : Macromolécule de poids moléculaire élevé, formée de polymères de nucléotides dont le sucre est le 2-désoxyribose. Il se présente sous forme d'une double chaîne hélicoïdale dont les deux brins sont complémentaires. Il constitue le génome de la plupart des organismes vivants.

ADN complémentaire : Ecrit aussi ADNc, simple brin artificiellement synthétisé à partir d'un ARNm, représente ainsi la partie codante de la région du génome qui a été transcrite en cet ARNm. Il est obtenu après une réaction de transcription inverse d'un ARNm mature.

Alignement de séquences : Manière de disposer les composantes des ADN, ARN ou acides aminés pour identifier les zones de concordances qui traduisent des similarités ou dissemblances. Des gaps peuvent être inclus dans l'alignement pour aligner les caractères communs sur des colonnes successives.

Apoptose : représente la mort cellulaire génétiquement régulée qui a lieu dans des cellules métaboliquement actives et qui est contrôlée par des mécanismes d'induction.

ARN : Macromolécule formée par la polymérisation de nombreux nucléotides dont le sucre est le ribose, présente dans le cytoplasme, les mitochondries ainsi que dans le noyau cellulaire, et servant d'intermédiaire dans la synthèse des protéines.

ARN messenger (ARNm) : Copie transitoire d'une portion de l'ADN correspondant à un ou plusieurs gènes utilisé comme intermédiaire par les cellules pour la synthèse des protéines.

Biopuce ADN : Ensemble de molécules d'ADN fixées en rangées ordonnées sur une petite surface permettant d'analyser le niveau d'expression de gènes (transcrits) d'un échantillon.

BLAST : Méthode de recherche heuristique utilisée en bio-informatique permettant de rechercher les régions similaires entre deux ou plusieurs séquences de nucléotides ou d'acides aminés.

Codon : Triplet de nucléotides A, C, U ou G de l'ARN messenger.

Cytoplasme : Désigne le contenu d'une cellule vivante. Plus exactement, il s'agit de la totalité du matériel cellulaire du protoplasme délimité par la membrane plasmique..

Électrophorèse : Technique utilisée en biologie pour la séparation et la caractérisation des molécules.

Épissage : Chez les eucaryotes, processus par lequel les ARN transcrits à partir de l'ADN génomique peuvent subir des étapes de coupure et ligature qui conduisent à la suppression de certaines régions dans l'ARNm. Les segments conservés s'appellent des exons et ceux qui sont éliminés s'appellent des introns.

Eucaryote : Organismes dont les cellules possèdent un noyau.

Exon : Parties transcrites des gènes qui codent des protéines.

Read : *mot* en anglais, courte séquence d'environ 35 à 50pb à l'état brut après son séquençage, en séquence couleur au lieu de nucléotides pour ABI SOLID.

Gap : Un gap apparaît dans un alignement entre deux séquences pour représenter une délétion dans l'une d'elle ou une insertion dans l'autre au niveau d'une paire de bases donnée.

Génome : ensemble du matériel génétique d'un individu ou d'une espèce codé dans son ADN (à l'exception de certains virus dont le génome est porté par des molécules d'ARN).

Histone : Protéine autour de laquelle l'ADN s'enroule pour former les chromosomes.

Indexage : Permet de retrouver une information plus rapidement et facilement

Intergénique : Séquence d'ADN non transcrite séparant les gènes.

Intron : Portion d'un gène non codante éliminée lors de l'épissage.

Intronique : Se dit d'une région contenant un intron d'une séquence avant son épissage.

Maturation de l'ARNm : Étapes permettant le passage de l'ARN à l'ARNm vers l'établissement d'une protéine.

Métastase : Foyer secondaire d'une affection, suppuration et surtout cancer, disséminée par voie sanguine ou lymphatique à partir d'un foyer primitif

Nucléotide : Molécule organique composée d'une nucléobase, d'un pentose et de 1 à 3 groupements phosphates. Certains nucléotides forment la base de l'ADN et de l'ARN.

Nucléotides fluorescents : Nucléotides formant des séquences d'ADN complémentaires de la séquence d'un gène ou d'une partie du génome préparées in vitro et couplées à des fluorochromes (Colorisation pour reconnaissance visuelle).

Oligo : Un petit polymère de deux à vingt nucléotides.

PCR (Polymerase Chain Reaction) : Méthode de biologie moléculaire d'amplification génique in vitro, qui permet de copier en grand nombre une séquence d'ADN ou d'ARN connue à partir d'une faible quantité d'acide nucléique.

Pipeline : Méthode ou modèle d'exécution séquentielle de tâches dans lequel le résultat d'une tâche consiste en l'entrée de la ou des tâches subséquentes.

Polymérase : Enzymes qui ont pour rôle la synthèse d'un brin de polynucléotide (ADN ou ARN), le plus souvent en utilisant un brin complémentaire comme matrice et des nucléotides triphosphosphate (NTP ou dNTP) comme monomères

Primer : Courte séquence d'ARN ou d'ADN, complémentaire du début d'une matrice, servant de point de départ à la synthèse du brin complémentaire de cette dernière par une ADN polymérase.

Régulation post-transcriptionnelle : Phase de la régulation de l'expression des gènes comprenant tous les mécanismes affectant directement les molécules d'ARN produite lors de la transcription.

Ribosome : Complexes ribonucléoprotéiques (c'est-à-dire composés de protéines et d'ARN) présents dans les cellules eucaryotes et procaryotes. Leur fonction est de synthétiser les protéines en décodant l'information contenue dans l'ARN messenger.

Séquence : Suite des acides aminés d'un peptide (structure primaire) ou suite des nucléotides d'un ADN ou d'un ARN, ou d'un segment de ceux-ci. Par convention, une structure de nucléotides est écrite de l'extrémité 5' à 3'

SNP (Single-Nucleotide Polymorphism) : Désigne le polymorphisme d'un seul nucléotide, qui est la variation (polymorphisme) d'une seule paire de bases du génome, entre individus d'une même espèce.

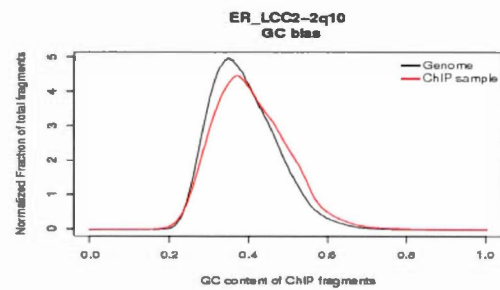
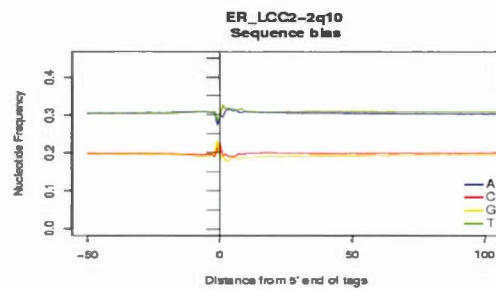
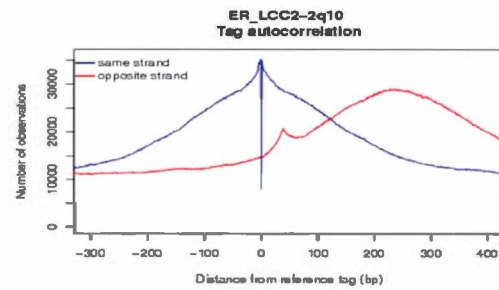
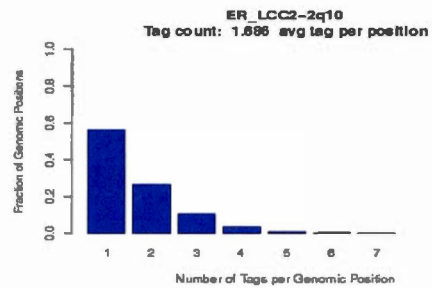
Synthèse protéique : Acte par lequel une cellule assemble une chaîne protéique en combinant des acides aminés isolés présents dans son cytoplasme, guidé par l'information contenue dans l'ADN. Elle se déroule en deux étapes au moins : la transcription de l'ADN en ARN messenger et la traduction de l'ARN messenger en une protéine.

Traduction : Interprétation des codons de l'ARNm en acides aminés.

Transcriptome : Ensemble des ARN issus de l'expression d'une partie du génome d'un tissu cellulaire ou d'un type de cellule.

Transformation Burrows-Wheeler : Technique utilisée en compression de données inventée par Michael Burrows et David Wheeler (Burrows & Wheeler 1994)

ANNEXE



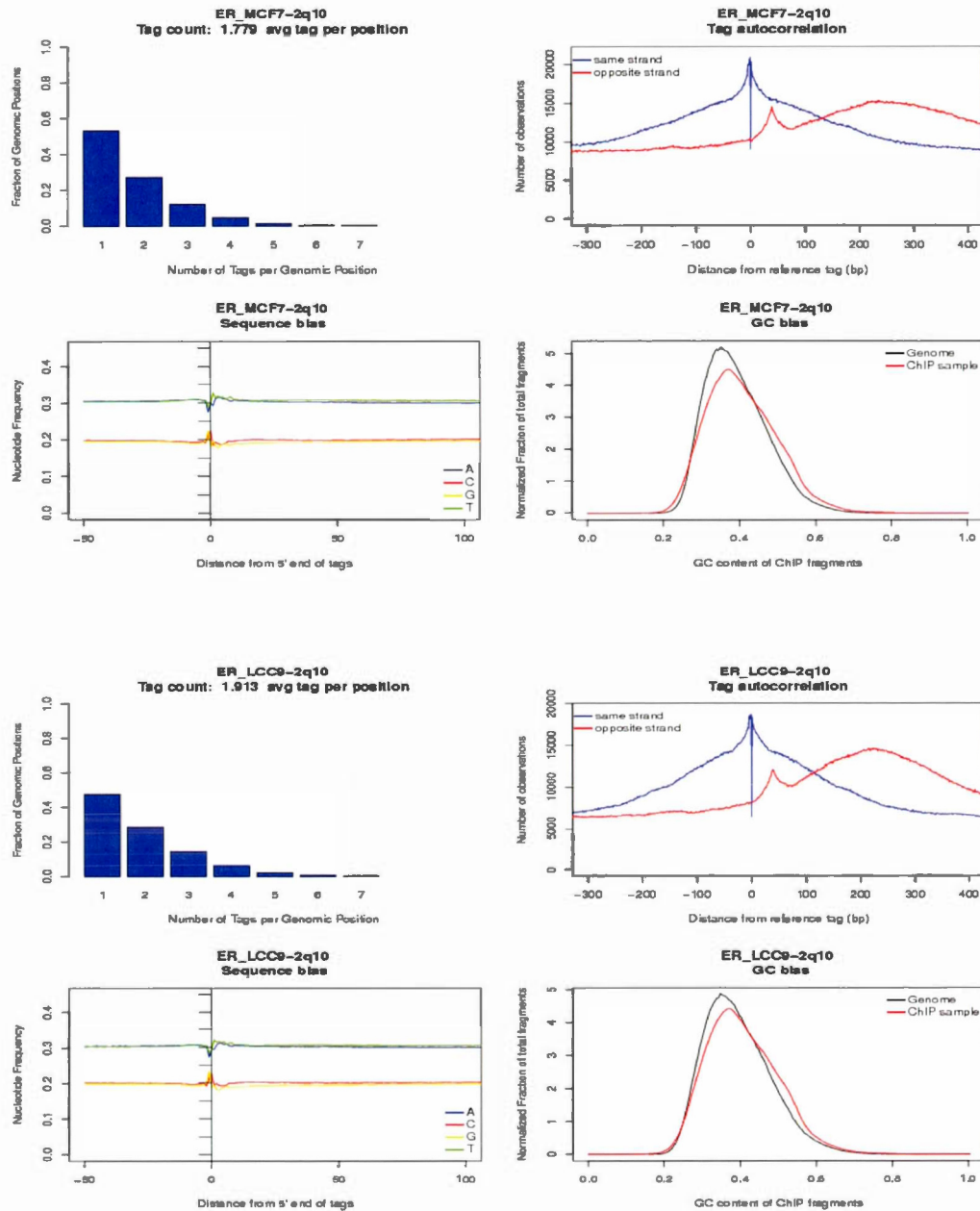
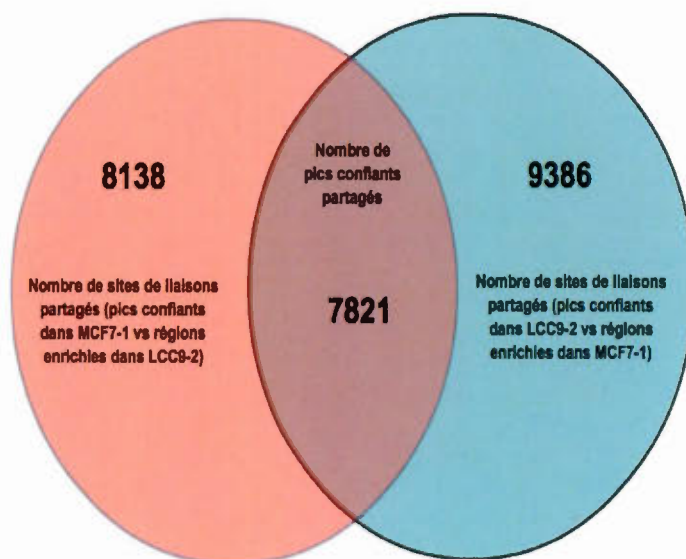
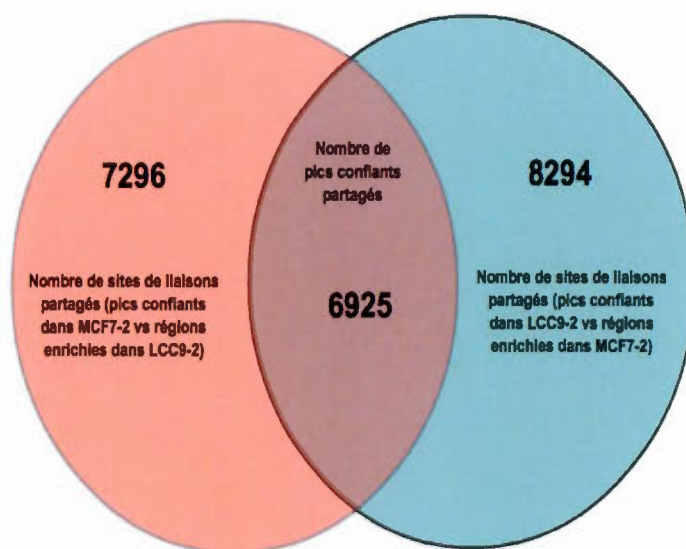
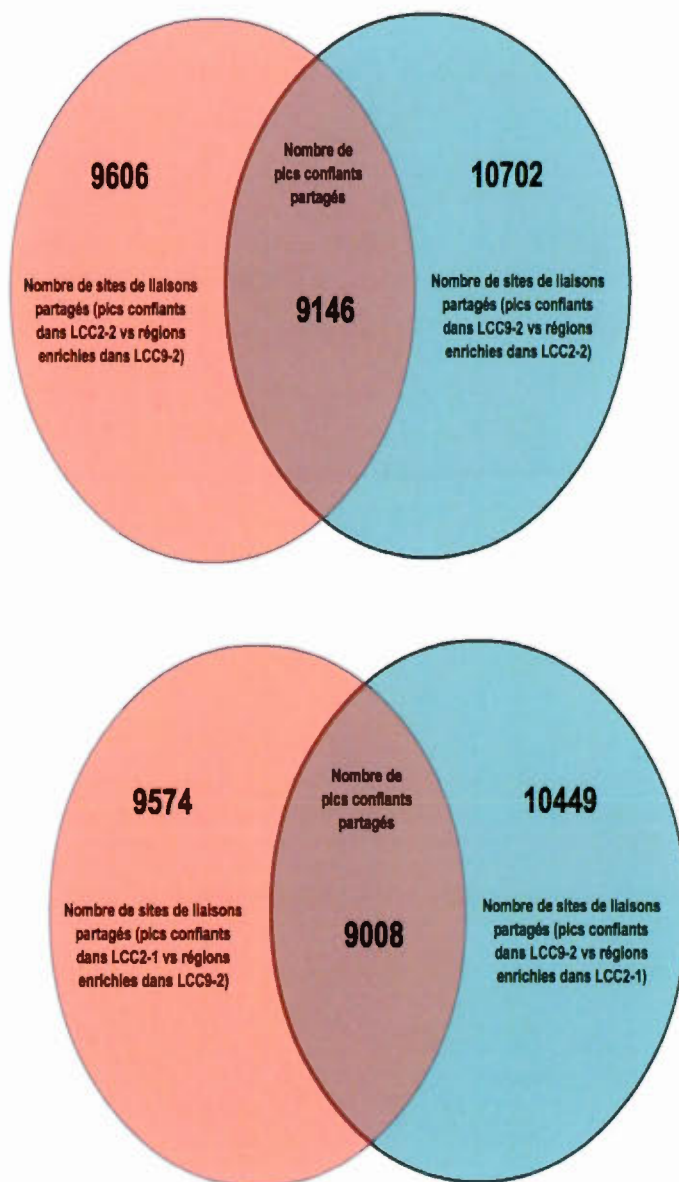
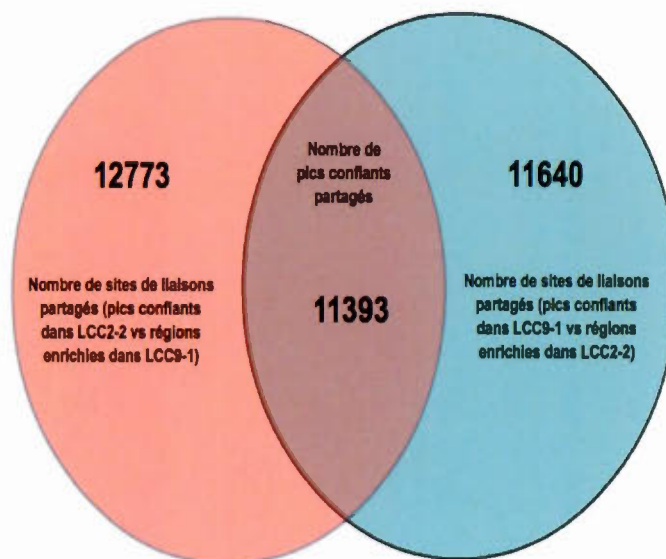
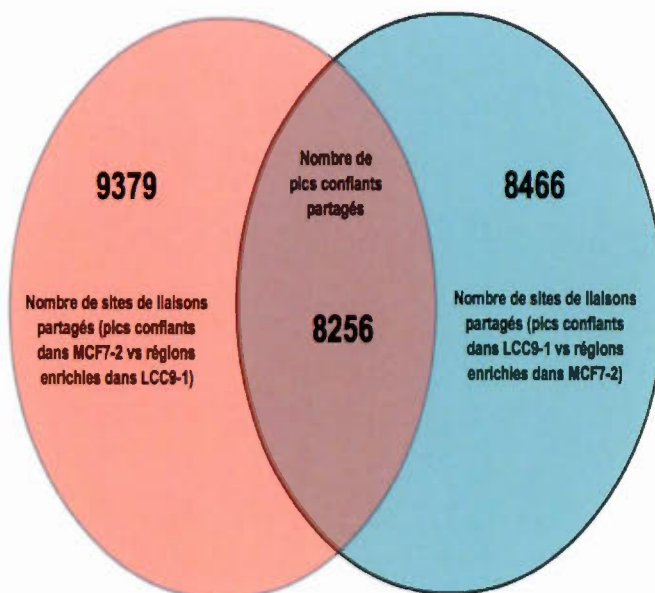
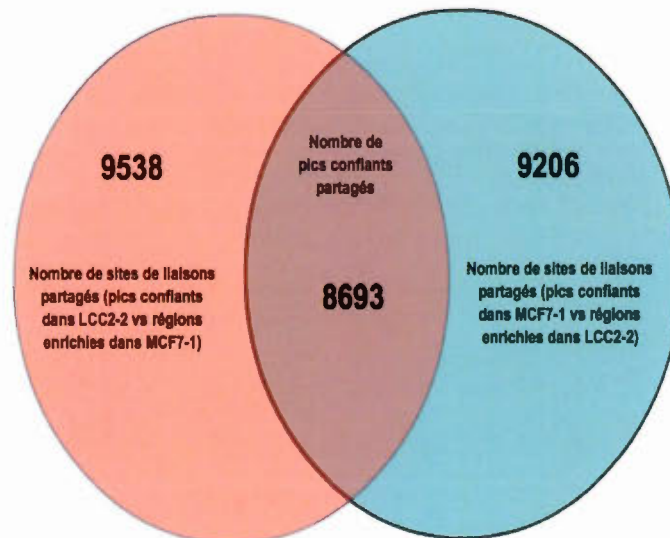
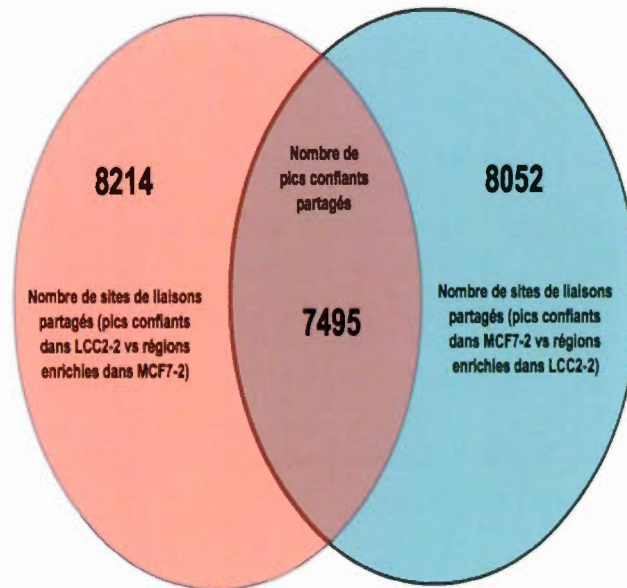


Figure S1: Les quatre mesures de qualité (le nombre de lecture par position unique, l'auto-correlation des "tag" le biais de séquence et le biais GC) obtenues à l'aide du logiciel HOMER dans chacune de nos trois expériences replicats pour nos trois expériences ChIP-Seq).









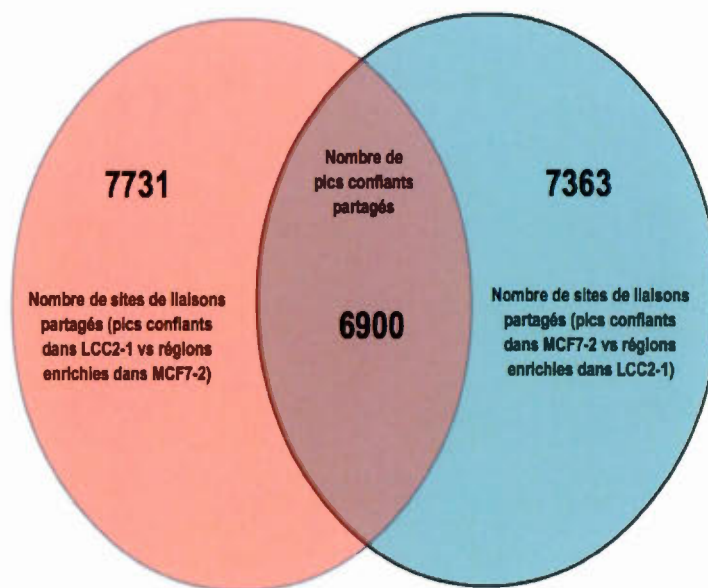
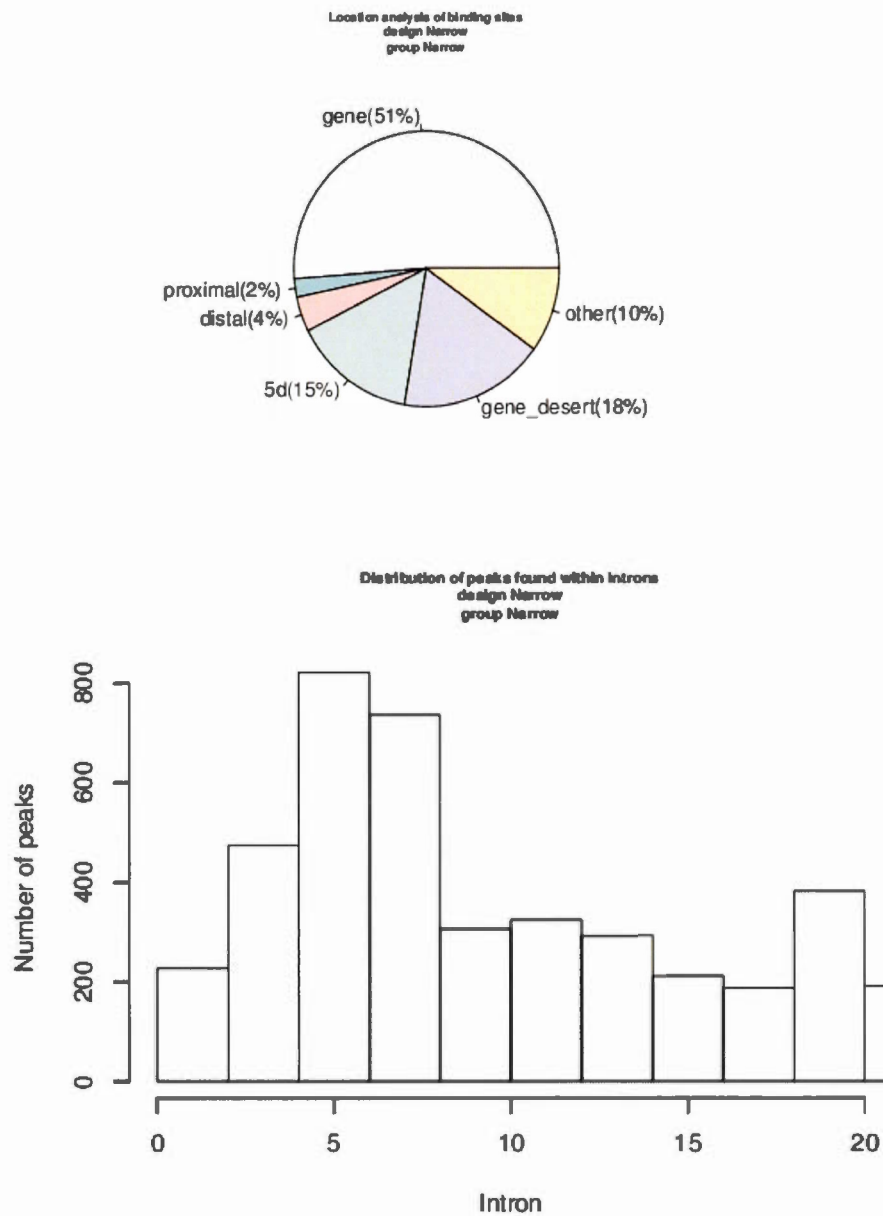
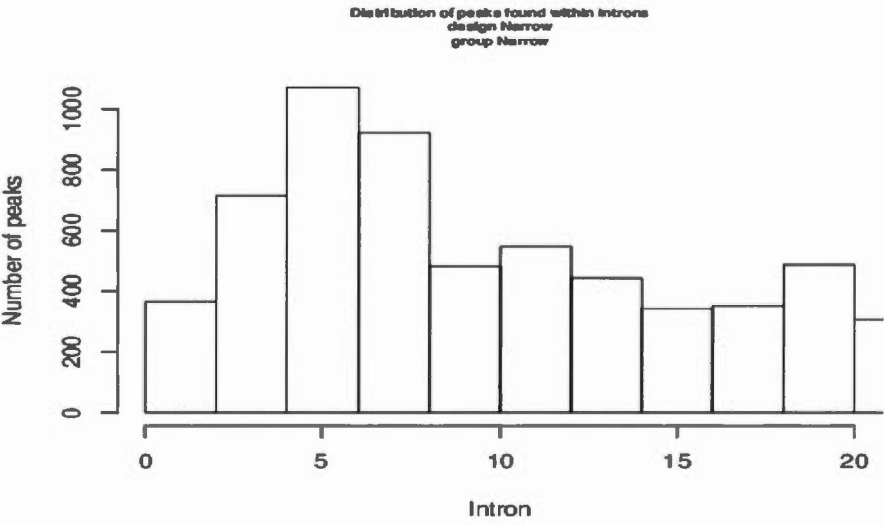
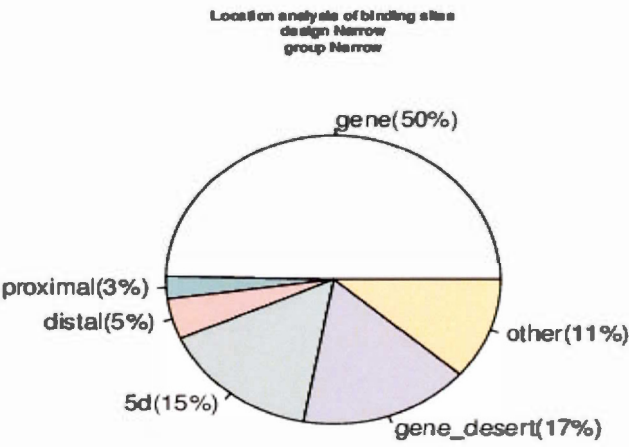
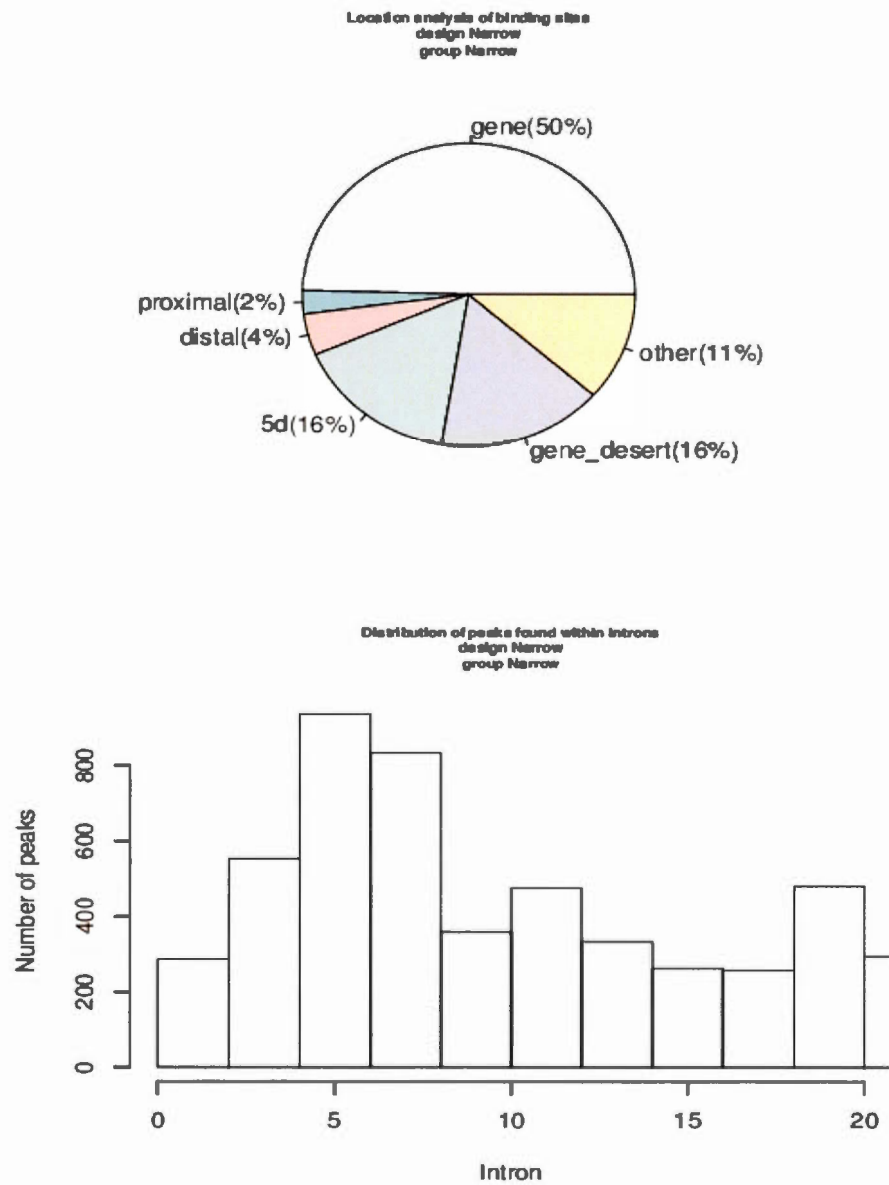


Figure S2: Le nombre de site de liaison partagés entre différentes combinaisons de réplicats deux à deux.

*MCF7-2*



LCC2-2



LCC9-2

Figure S3: Annotations des localisations des pics pour chacune des répliquats dans les trois conditions

Tableau S1: Exemple de fichiers de sortie de MACS

```

# This file is generated by MACS version 2.1.0.20140616
# Command line: callpeak --format BAM --fix-bimodal --gsize 2509729011.2 --treatment alignment/ER_LCC2-1q10/ER_
# ARGUMENTS LIST:
# name = peak_call/Narrow/Narrow
# format = BAM
# ChIP-seq file = ['alignment/ER_LCC2-1q10/ER_LCC2-1q10.sorted.dup.bam']
# control file = None
# effective genome size = 2.51e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 10000 bps
# Broad region calling is off

# tag size is determined as 40 bps
# total tags in treatment: 5409034
# tags after filtering in treatment: 5408789
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.00
# d = 249
# alternative fragment length(s) may be 249 bps
# local lambda is disabled!

```

Tableau S2: Matrice du motif ERE (LCC2-1 vs MCF7-1 increase)

```

>GGTCACNSTGAC 1-GGTCACNSTGAC, BestGuess:ERE(NR), IR3/MCF7-ERa-ChIP-Seq(Unpublished)/Homer(0.935)      8.007313      -3185.470051      0
T:2358.0(12.25%), B:423.0(1.38%), P:1e-1383
0.052  0.007  0.825  0.116
0.025  0.014  0.938  0.023
0.039  0.028  0.084  0.849
0.013  0.883  0.055  0.049
0.874  0.017  0.049  0.060
0.104  0.431  0.261  0.205
0.224  0.289  0.239  0.248
0.186  0.277  0.404  0.133
0.073  0.049  0.018  0.860
0.049  0.042  0.899  0.010
0.819  0.094  0.042  0.045
0.022  0.928  0.018  0.032

```

Tableau S3: Matrice du motif FOXA1 (LCC2-1 vs MCF7-1 increase)

```

>TGTTTRCTYW 3-TGTTTRCTYW, BestGuess:FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer(0.951)      7.058687      -1366.729685      0
T:5369.0(27.89%), B:4148.2(13.53%), P:1e-593
0.105  0.001  0.053  0.841
0.296  0.001  0.702  0.001
0.001  0.001  0.001  0.997
0.001  0.001  0.001  0.997
0.001  0.001  0.112  0.886
0.485  0.001  0.513  0.001
0.052  0.863  0.001  0.084
0.262  0.247  0.001  0.490
0.158  0.367  0.131  0.345
0.485  0.001  0.068  0.446

```

Tableau S4: Matrice du motif AP-2g (LCC2-1 vs MCF7-1 increase)

```

>DSCCYSGGSA 3-DSCCYSGGSA, BestGuess:AP-2gamma (AP2) /MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer(0.938) 6.417419 -1161.812102 0
T:4492.0(23.34%),B:3396.9(11.08%),P:1e-504
0.226 0.061 0.357 0.356
0.045 0.439 0.446 0.070
0.104 0.861 0.013 0.022
0.005 0.645 0.001 0.349
0.126 0.353 0.089 0.433
0.218 0.369 0.292 0.121
0.568 0.001 0.281 0.150
0.097 0.001 0.901 0.001
0.004 0.001 0.994 0.001
0.075 0.471 0.339 0.115
0.373 0.334 0.075 0.217
0.414 0.164 0.232 0.191

```

Tableau S5: Matrice du motif GATA3 (LCC2-1 vs MCF7-1 increase)

```

>TTATCTGATYWK 14-TTATCTGATYWK, BestGuess:GATA3 (Zf) /iTreg-Gata3-ChIP-Seq(GSE20898)/Homer(0.798) 9.008276 -162.011018 0
T:629.0(3.27%),B:452.8(1.48%),P:1e-70
0.111 0.016 0.003 0.870
0.058 0.020 0.009 0.913
0.972 0.010 0.009 0.009
0.010 0.003 0.007 0.980
0.001 0.893 0.033 0.073
0.070 0.001 0.001 0.928
0.046 0.079 0.748 0.127
0.815 0.050 0.027 0.108
0.106 0.003 0.014 0.877
0.211 0.373 0.077 0.340
0.440 0.001 0.086 0.473
0.231 0.078 0.390 0.301

```


Tableau 2S: Go term

LCC2-1_vs_LCC9-1					
Inavariant			Increase		
Term	Enrichment	Genes in term	Term	Enrichment	Genes in term
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-diffuse_large_B_cell_lymphoma	9.42E-33	3817	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	3.21E-73	7849
diffuse_large_B_cell_lymphoma	9.51E-33	4139	central_nervous_system-brain-primitive_neuroectodermal_tumour	2.10E-70	7735
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Burkitt_lymphoma	1.42E-29	2246	carcinoid-endocrine_tumour	1.09E-66	5698
Burkitt_lymphoma	1.42E-29	2246	pancreas-carcinoid-endocrine_tumour	1.33E-63	5194
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	1.51E-29	7849	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-diffuse_large_B_cell_lymphoma	5.13E-62	3817
carcinoid-endocrine_tumour	1.20E-28	5698	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Burkitt_lymphoma	5.71E-61	2246
pancreas-carcinoid-endocrine_tumour	6.49E-28	5194	Burkitt_lymphoma	5.71E-61	2246
central_nervous_system-brain-primitive_neuroectodermal_tumour-medulloblastoma	1.35E-21	7735	diffuse_large_B_cell_lymphoma	1.50E-59	4139
haematopoietic_neoplasm	2.83E-21	4976	kidney-other-neoplasm	2.32E-52	11174
lymphoid_neoplasm	1.61E-19	12110	haematopoietic_and_lymphoid_tissue	3.53E-51	12901
developmental process	1.00E-08	4660	single-organism developmental process	2.30E-26	4613
single-organism developmental process	1.17E-08	4613	developmental process	4.99E-26	4660
multicellular organismal development	3.99E-08	4098	multicellular organismal development	3.32E-25	4098
anatomical structure morphogenesis	8.84E-08	1934	system development	2.49E-23	3502
anatomical structure development	2.38E-07	4068	anatomical structure development	4.20E-22	4068
cell differentiation	3.67E-07	2807	regulation of biological quality	8.38E-22	2647
system development	4.66E-07	3502	regulation of signaling	4.74E-21	2550
cellular developmental process	6.50E-07	2931	regulation of cell communication	5.84E-21	2558
generation of neurons	2.96E-06	1201	cell differentiation	8.35E-21	2807
regulation of cellular component movement	4.33E-06	573	single-organism process	4.95E-20	12139
protein binding	3.51E-05	8308	protein binding	7.64E-17	8308
potassium channel activity	9.22E-05	118	cytoskeletal protein binding	1.76E-10	732
potassium ion transmembrane transporter activity	0.0002118	138	enzyme binding	7.73E-10	1297
butyrate-CoA ligase activity	0.00038955	8	phospholipid binding	1.54E-09	283
phospholipid-translocating ATPase activity	0.00064871	15	actin binding	6.75E-08	377
nucleoside-triphosphatase regulator activity	0.00069215	315	protein kinase binding	1.29E-07	419
RNA polymerase II transcription corepressor activity	0.00080004	23	lipid binding	2.59E-07	566
cytoskeletal protein binding	0.00081471	732	kinase binding	5.01E-07	466
RNA polymerase II transcription factor binding transcription factor activity involved in negative regulation of transcription	0.00092686	32	protein dimerization activity	1.68E-06	1018
GTPase regulator activity	0.0012627	298	guanyl-nucleotide exchange factor activity	4.59E-06	187
AACTT1_UNKNOWN	9.22E-22	1805	GOZG1T_ESR1_TARGETS_DN	1.61E-51	697
MASSARWEH_TAMOXIFEN_RESISTANCE_DN	1.26E-15	238	BHAT_ESR1_TARGETS_VIA_AKT1_UP	8.43E-48	272
GOZG1T_ESR1_TARGETS_DN	3.79E-14	697	SMID_BREAST_CANCER_BASAL_DN	5.43E-47	659
SMID_BREAST_CANCER_BASAL_DN	9.75E-14	659	AACTT1_UNKNOWN	7.05E-43	1805
DUTERTRE ESTRADIOL_RESPONSE_24HR_DN	8.55E-13	491	BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	6.59E-42	207
chr8q23	5.55E-12	22	DACOSTA_UV_RESPONSE_VIA_ERCC3_DN	3.30E-38	832
DACOSTA_UV_RESPONSE_VIA_ERCC3_DN	7.20E-12	832	DUTERTRE ESTRADIOL_RESPONSE_24HR_DN	2.31E-34	491
CTTTGT_VSLEF1_Q2	2.22E-10	1874	MASSARWEH_TAMOXIFEN_RESISTANCE_DN	7.81E-31	238
chr17q23	2.84E-10	64	MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	2.04E-30	1044
DACOSTA_UV_RESPONSE_VIA_ERCC3_COMMON_DN	3.25E-10	474	CUI_TCF21_TARGETS_2_DN	1.16E-28	804
synapse	6.12E-05	542	cytoplasm	3.41E-27	9661
cell junction	2.24E-04	821	cytoplasmic part	1.73E-14	7140
fascia adherens	3.01E-04	13	cell junction	1.90E-13	821
cytoplasm	3.28E-04	9661	endomembrane system	3.19E-13	3283
intercalated disc	3.58E-04	45	intracellular	2.78E-12	12802
cytoplasmic part	4.75E-04	7140	intracellular part	4.31E-12	12675
Golgi apparatus	6.13E-04	1283	plasma membrane part	1.27E-11	2144
cell-cell contact zone	9.36E-04	51	cell projection	1.28E-10	1401
perinuclear region of cytoplasm	1.07E-03	535	vesicle	3.60E-10	3185
sarcolemma	1.08E-03	96	cytoplasmic vesicle	1.00E-09	1051

LCC2-2_vs_LCC9-2					
Invariant			Increase		
Term	Enrichment	Genes in term	Term	Enrichment	Genes in term
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	2.61E-15	7849	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	1.82E-63	7849
diffuse_large_B_cell_lymphoma	2.89E-15	4139	central_nervous_system-brain-primitive_neuroectodermal_tur	4.08E-62	7735
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-diffuse_large_B_cell_lymphoma	3.80E-15	3817	Burkitt_lymphoma	2.97E-54	2246
Burkitt_lymphoma	1.80E-13	2246	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Bu	2.97E-54	2246
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Burkitt_lymphoma	1.80E-13	2246	carcinoid-endocrine_tumour	3.32E-54	5698
carcinoid-endocrine_tumour	1.81E-13	5698	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-dif	1.28E-53	3817
pancreas-carcinoid-endocrine_tumour	7.53E-12	5194	diffuse_large_B_cell_lymphoma	3.43E-52	4139
haematopoietic_neoplasm	1.45E-11	4976	pancreas-carcinoid-endocrine_tumour	7.68E-51	5194
prostate-carcinoma	2.82E-11	8442	kidney-other-neoplasm	6.27E-45	11174
central_nervous_system-brain-primitive_neuroectodermal_tumour-medulloblastoma	1.40E-10	7735	haematopoietic_and_lymphoid_tissue	9.25E-45	12901
localization of cell	1.13E-07	720	single-organism developmental process	2.75E-24	4613
cell motility	1.13E-07	720	developmental process	4.09E-24	4660
gland morphogenesis	1.46E-07	104	multicellular organismal development	1.40E-23	4098
cell migration	1.60E-07	659	system development	4.56E-22	3502
gland development	2.84E-07	374	regulation of biological quality	6.70E-22	2647
tissue morphogenesis	3.26E-07	473	cell differentiation	9.53E-22	2807
anatomical structure morphogenesis	3.88E-07	1934	anatomical structure development	1.39E-20	4068
anatomical structure development	4.08E-07	4068	cellular developmental process	1.98E-20	2931
developmental process	4.15E-07	4660	regulation of signaling	1.29E-18	2550
multicellular organismal development	5.94E-07	4098	regulation of cell communication	2.09E-18	2558
ion binding	1.85E-04	5942	protein binding	6.37E-14	8308
vascular endothelial growth factor receptor binding	3.52E-04	9	phospholipid binding	3.45E-08	283
Wnt-activated receptor activity	0.00035389	21	protein domain specific binding	6.10E-08	561
receptor regulator activity	0.00049373	40	cytoskeletal protein binding	1.08E-07	732
Wnt-protein binding	0.00110528	28	enzyme binding	1.40E-07	1297
metal ion binding	0.00111454	4015	sequence-specific DNA binding RNA polymerase II transcription	3.08E-07	337
cation binding	0.00126753	4086	lipid binding	1.38E-06	566
phosphotransferase activity, alcohol group as acceptor	0.00156595	705	regulatory region DNA binding	1.39E-06	481
PH domain binding	0.00160742	4	regulatory region nucleic acid binding	1.39E-06	481
kinase activity	0.00208645	765	transcription regulatory region DNA binding	2.29E-06	474
GOZGIT_ESR1_TARGETS_DN	1.48E-12	697	BHAT_ESR1_TARGETS_VIA_AKT1_UP	2.18E-58	272
chr17q23	3.38E-12	64	BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	2.27E-53	207
XU_GH1_AUTOCRINE_TARGETS_DN	1.09E-08	134	GOZGIT_ESR1_TARGETS_DN	1.34E-52	697
NIKOLSKY_BREAST_CANCER_17Q21_Q25_AMPLICON	1.31E-08	327	SMID_BREAST_CANCER_BASAL_DN	2.55E-51	659
NIKOLSKY_BREAST_CANCER_20Q12_Q13_AMPLICON	8.17E-08	134	AACITTT_UNKNOWN	3.88E-40	1805
chr1p13	1.08E-07	97	MASSARWEH_TAMOXIFEN_RESISTANCE_DN	1.23E-33	238
SMID_BREAST_CANCER_BASAL_DN	1.40E-07	659	DACOSTA_UV_RESPONSE_VIA_ERCC3_DN	1.76E-33	832
MASSARWEH_TAMOXIFEN_RESISTANCE_DN	1.76E-07	238	MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	1.14E-30	1044
V5TAL1ALPHA47_01	8.18E-07	237	RAF_UP.V1_DN	1.24E-30	188
CHARAFE_BREAST_CANCER_LUMINAL_V5_MESENCHYMAL_UP	1.55E-06	427	CUI_TCF21_TARGETS_2_DN	5.28E-30	804
postsynaptic density	0.00016545	118	cytoplasm	1.81E-23	9661
synaptic vesicle membrane	0.00019965	51	cell junction	8.26E-13	821
laminin-2 complex	0.00027918	2	plasma membrane part	6.41E-12	2144
rough endoplasmic reticulum	0.00241354	56	cytoplasmic part	8.26E-12	7140
synapse part	0.00254487	391	endomembrane system	4.67E-11	3283
rough endoplasmic reticulum membrane	0.00265148	17	cell projection	1.23E-10	1401
extracellular matrix	0.00280182	395	intracellular part	7.42E-10	12675
cell cortex	0.00300742	211	intracellular	1.14E-09	12802
dendritic spine	0.00339317	87	vesicle	2.00E-09	3185
neuron spine	0.00359145	88	cell projection part	1.11E-08	654

LCC2-2_vs_MCF7-2					
Invariant			Increase		
Term	Enrichment	Genes in term	Term	Enrichment	Genes in term
carcinoid-endocrine_tumour	1.54E-11	5698	central_nervous_system-brain-primitive_neuroectodermal_tur	1.82E-63	7735
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	4.16E-11	7849	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	6.88E-58	7849
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-diffuse_large_B_cell_lymphom	4.33E-10	3817	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Bu	3.69E-53	2246
pancreas-carcinoid-endocrine_tumour	9.13E-10	5194	Burkitt_lymphoma	3.69E-53	2246
diffuse_large_B_cell_lymphoma	1.29E-08	4139	carcinoid-endocrine_tumour	4.03E-53	5698
haematopoietic_neoplasm	4.14E-08	4976	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-dif	6.06E-53	3817
large_intestine-caecum-carcinoma-adenocarcinoma	4.21E-08	14592	diffuse_large_B_cell_lymphoma	7.47E-53	4139
prostate-carcinoma	4.21E-08	8442	pancreas-carcinoid-endocrine_tumour	9.36E-51	5194
caecum	4.90E-08	14607	kidney-other-neoplasm	2.03E-46	11174
lung-carcinoma-small_cell_carcinoma	6.23E-08	10627	prostate-carcinoma	4.61E-44	8442
protein dephosphorylation	2.35E-04	156	multicellular organismal development	7.04E-22	4098
cellular component organization	8.04E-04	4282	single-organism developmental process	3.55E-21	4613
cell cycle	8.11E-04	1228	developmental process	8.40E-21	4660
cell cycle process	8.33E-04	952	system development	1.14E-19	3502
positive regulation of smooth muscle cell migration	9.26E-04	15	regulation of biological quality	3.59E-19	2647
cellular response to lipoprotein particle stimulus	1.03E-03	4	anatomical structure development	4.55E-18	4068
cellular component organization or biogenesis	1.07E-03	4391	anatomical structure morphogenesis	6.98E-17	1934
response to drug	1.58E-03	377	cell differentiation	5.80E-16	2807
trigeminal nerve development	1.70E-03	5	localization	1.06E-15	4266
positive regulation of melanocyte differentiation	1.70E-03	5	tissue development	1.25E-15	1430
RNA polymerase II transcription corepressor activity	0.00019533	23	protein binding	7.07E-13	8308
co-SMAD binding	0.00042527	12	phospholipid binding	4.12E-10	283
RNA polymerase II transcription factor binding transcription factor activity involved in nega	0.00072455	32	cytoskeletal protein binding	1.17E-08	732
phosphoprotein phosphatase activity	0.00157286	169	lipid binding	1.90E-07	566
protein tyrosine phosphatase activity	0.00215369	103	enzyme binding	2.53E-07	1297
RNA polymerase II transcription factor binding transcription factor activity	0.00313421	111	beta-catenin binding	2.09E-06	62
RNA polymerase II transcription cofactor activity	0.00314895	77	Ras guanyl-nucleotide exchange factor activity	2.31E-06	120
repressing transcription factor binding	0.00446552	52	guanyl-nucleotide exchange factor activity	2.98E-06	187
voltage-gated potassium channel activity	0.00505886	86	phosphatidylinositol binding	6.26E-06	170
phosphatase activity	0.00649535	259	potassium channel activity	8.06E-06	118
NIKOLSKY_BREAST_CANCER_17Q21_Q25_AMPICON	5.76E-12	327	GOZGIT_ESR1_TARGETS_DN	2.61E-57	697
chr17q23	2.67E-11	64	BHAT_ESR1_TARGETS_VIA_AKT1_UP	2.50E-50	272
GOZGIT_ESR1_TARGETS_DN	2.36E-08	697	BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	4.91E-46	207
MASSARWEH_TAMOXIFEN_RESISTANCE_DN	7.61E-08	238	SMID_BREAST_CANCER_BASAL_DN	1.73E-45	659
RAF_UP.V1_DN	1.45E-07	188	AACTT1_UNKNOWN	6.42E-42	1805
NIKOLSKY_BREAST_CANCER_20Q12_Q13_AMPICON	1.56E-07	134	MASSARWEH_TAMOXIFEN_RESISTANCE_DN	6.68E-35	238
MODULE_331	3.78E-07	73	DACOSTA_UV_RESPONSE_VIA_ERCC3_DN	1.56E-32	832
FARMER_BREAST_CANCER_CLUSTER_6	1.28E-06	16	DUTERTRE ESTRADIOL_RESPONSE_24HR_DN	2.26E-32	491
chr1p13	4.26E-06	97	RAF_UP.V1_DN	5.42E-32	188
SMID_BREAST_CANCER_BASAL_DN	4.30E-06	659	CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP	4.72E-29	427
intracellular organelle	6.17E-04	11029	cytoplasm	7.77E-22	9661
intracellular membrane-bounded organelle	8.00E-04	9975	cell projection	1.84E-11	1401
organelle	2.69E-03	11972	cytoplasmic part	3.99E-11	7140
organelle part	2.74E-03	7152	plasma membrane part	7.77E-11	2144
platelet alpha granule membrane	3.20E-03	7	endomembrane system	8.13E-11	3283
cell junction	3.83E-03	821	intracellular	2.32E-10	12802
stereocilium	4.17E-03	26	intracellular part	3.23E-10	12675
cell part	5.21E-03	15016	plasma membrane region	4.54E-10	413
cell	5.28E-03	15018	cell junction	1.62E-09	821
endomembrane system	5.74E-03	3283	cell projection part	2.61E-09	654

LCC9-1_vs_MCF7-1					
Invariant			Increase		
Term	Enrichment	Genes in term	Term	Enrichment	Genes in term
central_nervous_system-brain-primitive_neuroectodermal_tumour-medulloblastoma	1.33E-13	7735	central_nervous_system-brain-primitive_neuroectodermal_tur	4.92E-76	7735
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-diffuse_large_B_cell_lymphoma	1.73E-10	3817	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	5.76E-76	7849
diffuse_large_B_cell_lymphoma	5.45E-10	4139	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-dif	3.72E-66	3817
pancreas-carcinoid-endocrine_tumour	2.02E-09	5194	diffuse_large_B_cell_lymphoma	5.15E-65	4139
carcinoid-endocrine_tumour	2.05E-09	5698	Burkitt_lymphoma	8.07E-64	2246
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	2.21E-08	7849	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Bu	8.07E-64	2246
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Burkitt_lymphoma	4.03E-08	2246	carcinoid-endocrine_tumour	4.55E-63	5698
Burkitt_lymphoma	4.03E-08	2246	pancreas-carcinoid-endocrine_tumour	2.54E-59	5194
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-chronic_lymphocytic_leukaem	7.89E-08	3552	kidney-other-neoplasm	7.30E-54	11174
chronic_lymphocytic_leukaemia-small_lymphocytic_lymphoma	8.01E-08	3553	single-organism developmental process	1.11E-27	4613
system development	1.28E-05	3502	developmental process	1.18E-27	4660
multicellular organismal development	2.59E-05	4098	multicellular organismal development	4.07E-25	4098
single-organism developmental process	2.93E-05	4613	anatomical structure development	1.99E-23	4068
cellular developmental process	3.73E-05	2931	system development	2.63E-22	3502
cell differentiation	4.00E-05	2807	cell differentiation	5.35E-22	2807
developmental process	4.84E-05	4660	cellular developmental process	3.01E-21	2931
reproductive structure development	5.34E-05	401	regulation of biological quality	3.59E-21	2647
reproductive system development	5.74E-05	403	anatomical structure morphogenesis	9.92E-20	1934
kidney development	6.32E-05	232	regulation of localization	4.16E-19	1784
positive regulation of cellular process	1.01E-04	3632	haematopoietic_and_lymphoid_tissue	3.53E-51	12901
RNA polymerase II core promoter proximal region sequence-specific DNA binding transcrip	0.00012753	165	protein binding	3.94E-18	8308
RNA polymerase II core promoter proximal region sequence-specific DNA binding transcrip	0.00016395	121	phospholipid binding	1.47E-09	283
binding	0.0001648	12517	enzyme binding	1.76E-09	1297
RNA polymerase II transcription regulatory region sequence-specific DNA binding transcrip	0.00026786	153	cytoskeletal protein binding	1.74E-08	732
sequence-specific DNA binding RNA polymerase II transcription factor activity	0.00048746	337	lipid binding	4.19E-08	566
protein binding, bridging	0.00131837	131	actin binding	4.88E-07	377
clathrin binding	0.0023523	28	protein dimerization activity	1.67E-06	1018
binding, bridging	0.00230087	142	binding	2.35E-06	12517
clathrin adaptor activity	0.00234987	4	protein kinase binding	2.69E-06	419
transcription regulatory region DNA binding	0.00344692	474	kinase binding	3.51E-06	466
chr1p13	6.61E-13	97	GOZGIT_ESR1_TARGETS_DN	7.10E-61	697
GOZGIT_ESR1_TARGETS_DN	1.34E-11	697	BHAT_ESR1_TARGETS_VIA_AKT1_UP	9.41E-49	272
AACITTT_UNKNOWN	7.22E-09	1805	BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	3.15E-46	207
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	2.11E-08	1044	SMID_BREAST_CANCER_BASAL_DN	9.71E-46	659
SMID_BREAST_CANCER_BASAL_DN	4.81E-08	659	AACITTT_UNKNOWN	5.50E-45	1805
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_UP	8.50E-07	356	DUTERTRE ESTRADIOL_RESPONSE_24HR_DN	1.62E-34	491
AAGCAC/MIR-218	1.06E-06	388	MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	4.02E-32	1044
RAF_UP.V1_DN	1.22E-06	188	DACOSTA_UV_RESPONSE_VIA_ERCC3_DN	4.47E-32	832
MODULE_331	1.80E-06	73	RAF_UP.V1_DN	4.97E-32	188
CTTGAT/MIR-381	2.58E-06	199	CUI_TCF21_TARGETS_2_DN	1.96E-30	804
dendritic spine	8.65E-06	87	cytoplasm	1.31E-26	9661
neuron spine	9.59E-06	88	cell junction	1.80E-15	821
cell junction	2.22E-04	821	cytoplasmic part	2.15E-14	7140
synapse	2.25E-04	542	plasma membrane part	6.27E-13	2144
dendritic shaft	5.21E-04	34	endomembrane system	8.60E-13	3283
dendrite	7.58E-04	354	vesicle	1.08E-12	3185
excitatory synapse	4.30E-03	17	cytoplasmic vesicle	1.34E-11	1051
membrane-bounded organelle	4.87E-03	11184	membrane-bounded vesicle	1.37E-11	3107
COP9 signalosome	4.93E-03	35	intracellular part	3.58E-11	12675
pi-body	5.60E-03	6	intracellular	6.89E-11	12802

LCC9-2_vs_MCF7-2					
Invariant			Increase		
Term	Enrichment	Genes in term	Term	Enrichment	Genes in term
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-diffuse_large_B_cell_lymphoma	1.76E-13	3817	central_nervous_system-brain-primitive_neuroectodermal_tumour	4.34E-61	7735
diffuse_large_B_cell_lymphoma	5.50E-12	4139	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	1.29E-56	7849
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm	1.04E-10	7849	carcinoid-endocrine_tumour	5.52E-53	5698
carcinoid-endocrine_tumour	1.08E-10	5698	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-dif	3.54E-52	3817
kidney-other-neoplasm	1.96E-10	11174	diffuse_large_B_cell_lymphoma	2.42E-51	4139
pancreas-carcinoid-endocrine_tumour	5.73E-10	5194	haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Bu	1.14E-50	2246
central_nervous_system-brain-primitive_neuroectodermal_tumour-medulloblastoma	1.52E-09	7735	Burkitt_lymphoma	1.14E-50	2246
Burkitt_lymphoma	2.53E-09	2246	pancreas-carcinoid-endocrine_tumour	1.02E-49	5194
haematopoietic_and_lymphoid_tissue-lymphoid_neoplasm-Burkitt_lymphoma	2.53E-09	2246	prostate-carcinoma	3.25E-43	8442
haematopoietic_neoplasm	8.84E-09	4976	kidney-other-neoplasm	4.00E-43	11174
gland morphogenesis	2.56E-06	104	multicellular organismal development	5.39E-19	4098
epithelial cell differentiation	8.10E-06	448	single-organism developmental process	5.94E-19	4613
positive regulation of neuroepithelial cell differentiation	2.90E-05	5	developmental process	7.50E-19	4660
positive regulation of melanocyte differentiation	2.90E-05	5	regulation of biological quality	2.13E-18	2647
epithelium development	4.98E-05	909	cell differentiation	4.01E-17	2807
urogenital system development	5.55E-05	279	system development	1.47E-16	3502
renal system development	5.57E-05	244	anatomical structure development	1.68E-15	4068
columnar/cuboidal epithelial cell differentiation	6.56E-05	70	cellular developmental process	2.21E-15	2931
regulation of extrinsic apoptotic signaling pathway	8.03E-05	154	localization	2.39E-14	4266
limb development	8.48E-05	155	regulation of cell communication	3.70E-14	2558
inositol tetrakisphosphate kinase activity	0.00020498	2	protein binding	1.06E-10	8308
1-phosphatidylinositol binding	0.00320307	21	protein domain specific binding	1.61E-06	561
inositol trisphosphate kinase activity	0.00542145	8	phospholipid binding	2.60E-06	283
ion binding	0.00618692	5942	cytoskeletal protein binding	3.27E-06	732
metal ion binding	0.0083628	4015	enzyme binding	3.37E-06	1297
sodium:amino acid symporter activity	0.00854982	10	potassium ion transmembrane transporter activity	4.07E-06	138
chaperone binding	0.00900206	57	potassium channel activity	4.42E-06	118
amide transmembrane transporter activity	0.01035161	11	transmembrane transporter activity	4.54E-06	920
channel regulator activity	0.01187005	95	GTPase regulator activity	4.84E-06	298
cation:amino acid symporter activity	0.01230538	12	voltage-gated potassium channel activity	5.57E-06	86
NIKOLSKY_BREAST_CANCER_20Q12_Q13_AMPLICON	7.29E-10	134	GOZGIT_ESR1_TARGETS_DN	4.90E-57	697
GOZGIT_ESR1_TARGETS_DN	1.17E-09	697	BHAT_ESR1_TARGETS_VIA_AKT1_UP	2.34E-51	272
MASSARWEH_TAMOXIFEN_RESISTANCE_DN	1.31E-09	238	BHAT_ESR1_TARGETS_NOT_VIA_AKT1_UP	7.75E-51	207
AACTT_UNKNOWN	1.60E-09	1805	SMID_BREAST_CANCER_BASAL_DN	4.67E-49	659
NIKOLSKY_BREAST_CANCER_17Q21_Q25_AMPLICON	1.61E-09	327	AACTT_UNKNOWN	3.42E-36	1805
CTGTAT_MIR-381	3.57E-09	199	MASSARWEH_TAMOXIFEN_RESISTANCE_DN	3.94E-36	238
SMID_BREAST_CANCER_BASAL_DN	1.73E-08	659	RAF_UP.V1_DN	4.28E-35	188
CAGGTG_VSE12_Q6	7.54E-08	2375	CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_4	3.36E-29	283
STEIN_ESTROGEN_RESPONSE_NOT_VIA_ESRRA	1.21E-07	18	DUTERTRE ESTRADIOL_RESPONSE_6HR_UP	9.23E-28	222
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_4	1.21E-07	283	DUTERTRE ESTRADIOL_RESPONSE_24HR_DN	1.04E-27	491
intercellular bridge	1.20E-03	34	cytoplasm	2.54E-19	9661
lateral loop	1.86E-03	5	cell junction	9.56E-11	821
myelin sheath	2.01E-03	39	plasma membrane part	1.12E-10	2144
cell junction	2.39E-03	821	cytoplasmic part	1.55E-10	7140
asymmetric synapse	5.06E-03	8	endomembrane system	5.35E-10	3283
plasma membrane region	5.11E-03	413	cell projection	1.53E-09	1401
synaptic vesicle membrane	5.38E-03	51	cell projection part	6.54E-09	654
apical part of cell	1.09E-02	308	anchoring junction	1.32E-08	226
Golgi apparatus	1.11E-02	1283	cell periphery	2.38E-08	4650
Schmidt-Lanterman incisure	1.15E-02	12	plasma membrane	3.38E-08	4556

Tableau 3S: KEGG

LCC2-1_vs_LCC9-1												
Decrease				Invariant				Increase				
Term id	Term	Enrichment	Genes in term	Term id	Term	Enrichment	Genes in term	Term id	Term	Enrichment	Genes in term	
KEGG pathways	ko04360	Axon guidance	1.44E-04	127	ko04360	Axon guidance	1.44E-04	127	ko04360	Axon guidance	1.46E-08	127
	hsa04360	Axon guidance	1.44E-04	127	hsa04360	Axon guidance	1.44E-04	127	hsa04360	Axon guidance	1.46E-08	127
	hsa00650	Butanoate met	1.34E-03	26	hsa00650	Butanoate met	1.34E-03	26	hsa05200	Pathways in ca	6.00E-08	327
	ko00650	Butanoate met	1.34E-03	26	ko00650	Butanoate met	1.34E-03	26	hsa05205	Proteoglycans i	2.34E-06	225
	ko04540	Gap junction	1.16E-02	89	ko04540	Gap junction	1.16E-02	89	ko05205	Proteoglycans i	2.34E-06	225
	hsa04540	Gap junction	1.16E-02	89	hsa04540	Gap junction	1.16E-02	89	hsa04015	Rap1 signaling	2.49E-06	211
	hsa05205	Proteoglycans i	1.45E-02	225	hsa05205	Proteoglycans i	1.45E-02	225	ko04015	Rap1 signaling	2.49E-06	211
	ko05205	Proteoglycans i	1.45E-02	225	ko05205	Proteoglycans i	1.45E-02	225	hsa04510	Focal adhesion	3.74E-06	207
	hsa04390	Hippo signaling	1.79E-02	153	hsa04390	Hippo signaling	1.79E-02	153	ko04510	Focal adhesion	3.74E-06	207
	ko04390	Hippo signaling	1.79E-02	153	ko04390	Hippo signaling	1.79E-02	153	hsa04012	ErbB signaling	2.47E-05	87
LCC2-2_vs_LCC9-2												
Decrease				Invariant				Increase				
Term id	Term	Enrichment	Genes in term	Term id	Term	Enrichment	Genes in term	Term id	Term	Enrichment	Genes in term	
KEGG pathways	ko04360	Axon guidance	1.16E-03	127	ko04360	Axon guidance	5.28E-04	127	hsa04360	Axon guidance	2.69E-06	127
	hsa04360	Axon guidance	1.16E-03	127	hsa04360	Axon guidance	5.28E-04	127	ko04360	Axon guidance	2.69E-06	127
	ko05206	MicroRNAs in c	1.26E-03	294	hsa00562	Inositol phosph	5.15E-03	61	hsa05200	Pathways in ca	4.97E-06	327
	hsa05206	MicroRNAs in c	1.26E-03	294	ko00562	Inositol phosph	5.15E-03	61	ko04015	Rap1 signaling	5.05E-06	211
	ko05221	Acute myeloid	5.97E-03	57	hsa05131	Shigellosis	5.15E-03	61	hsa04015	Rap1 signaling	5.05E-06	211
	hsa05221	Acute myeloid	5.97E-03	57	hsa04062	Chemokine sign	8.01E-03	189	hsa05205	Proteoglycans i	7.60E-05	225
	ko05211	Renal cell carci	9.99E-03	66	ko04062	Chemokine sign	8.01E-03	189	ko05205	Proteoglycans i	7.60E-05	225
	hsa05211	Renal cell carci	9.99E-03	66	ko05205	Proteoglycans i	8.44E-03	225	ko04390	Hippo signaling	3.15E-04	153
	hsa04066	HIF-1 signaling	1.14E-02	106	hsa05205	Proteoglycans i	8.44E-03	225	hsa04390	Hippo signaling	3.15E-04	153
	ko05220	Chronic myeloi	1.41E-02	73	hsa04664	Fc epsilon RI sig	9.19E-03	70	ko04510	Focal adhesion	6.01E-04	207
LCC2-1_vs_MCF7-1												
Decrease				Invariant				Increase				
Term id	Term	Enrichment	Genes in term	Term id	Term	Enrichment	Genes in term	Term id	Term	Enrichment	Genes in term	
KEGG pathways	ko00534	Glycosaminogly	9.56E-03	24	ko00534	Glycosaminogly	9.56E-03	24	hsa05200	Pathways in ca	1.46E-07	327
	hsa00534	Glycosaminogly	9.56E-03	24	hsa00534	Glycosaminogly	9.56E-03	24	ko04015	Rap1 signaling	1.06E-06	211
	M00078	Heparan sulfat	1.13E-02	9	hsa_M00078	Heparan sulfat	1.13E-02	9	hsa04015	Rap1 signaling	1.06E-06	211
	hsa_M00078	Heparan sulfat	1.13E-02	9	M00078	Heparan sulfat	1.13E-02	9	ko05205	Proteoglycans i	1.32E-06	225
	hsa04350	TGF-beta signa	1.63E-02	80	ko04350	TGF-beta signa	1.63E-02	80	hsa05205	Proteoglycans i	1.32E-06	225
	ko04350	TGF-beta signa	1.63E-02	80	hsa04350	TGF-beta signa	1.63E-02	80	hsa04360	Axon guidance	3.41E-06	127
	ko05130	Pathogenic Esc	1.88E-02	55	hsa05130	Pathogenic Esc	1.88E-02	55	ko04360	Axon guidance	3.41E-06	127
	hsa05130	Pathogenic Esc	1.88E-02	55	ko05130	Pathogenic Esc	1.88E-02	55	hsa04666	Fc gamma R-mv	3.42E-06	91
	hsa04670	Leukocyte tran	2.23E-02	118	hsa04670	Leukocyte tran	2.23E-02	118	ko04666	Fc gamma R-mv	3.42E-06	91
	ko04670	Leukocyte tran	2.23E-02	118	ko04670	Leukocyte tran	2.23E-02	118	ko04144	Endocytosis	2.54E-05	203

BIBLIOGRAPHIE

- Acevedo, Luis G, A Leonardo Iniguez, Heather L Holster, Xinmin Zhang, Roland Green, and Peggy J Farnham. 2007. "Genome-Scale ChIP-Chip Analysis Using 10,000 Human Cells." *BioTechniques* 43 (6). England: 791–97.
- Ahmadiyeh, Nasim, Mark M Pomerantz, Chiara Grisanzio, Paula Herman, Li Jia, Vanessa Almendro, Housheng Hansen He, et al. 2010. "8q24 Prostate, Breast, and Colon Cancer Risk Loci Show Tissue-Specific Long-Range Interaction with MYC." *Proceedings of the National Academy of Sciences of the United States of America* 107 (21). United States: 9742–46. doi:10.1073/pnas.0910668107.
- Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. *Molecular Biology of the Cell, Fourth Edition*. Garland Science. citeulike-article-id:691434.
- Alkan, Can, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, et al. 2009. "Personalized Copy Number and Segmental Duplication Maps Using next-Generation Sequencing." *Nature Genetics* 41 (10). United States: 1061–67. doi:10.1038/ng.437.
- AlQuraishi, Mohammed, and Harley H McAdams. 2011. "Direct Inference of Protein-DNA Interactions Using Compressed Sensing Methods." *Proceedings of the National Academy of Sciences of the United States of America* 108 (36). United States: 14819–24. doi:10.1073/pnas.1106460108.
- Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3). ENGLAND: 403–10. doi:10.1016/S0022-2836(05)80360-2.
- Anderson, Elizabeth. 2002. "The Role of Oestrogen and Progesterone Receptors in Human Mammary Development and Tumorigenesis." *Breast Cancer Research : BCR* 4 (5). England: 197–201.
- Ansorge, Wilhelm J. 2009. "Next-Generation DNA Sequencing Techniques." *New Biotechnology* 25 (4). Netherlands: 195–203. doi:10.1016/j.nbt.2008.12.009.
- Arteaga, Carlos L, Mark X Sliwkowski, C Kent Osborne, Edith A Perez, Fabio Puglisi, and Luca Gianni. 2012. "Treatment of HER2-Positive Breast Cancer: Current Status and Future Perspectives." *Nat Rev Clin Oncol* 9 (1). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 16–32. http://dx.doi.org/10.1038/nrclinonc.2011.177.
- Ascenzi, Paolo, Alessio Bocedi, and Maria Marino. 2006. "Structure-Function Relationship of Estrogen Receptor Alpha and Beta: Impact on Human Health." *Molecular Aspects of Medicine* 27 (4). England: 299–402. doi:10.1016/j.mam.2006.07.001.
- Bailey, Timothy, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. 2013. "Practical Guidelines for the Comprehensive

- Analysis of ChIP-Seq Data." *PLoS Computational Biology* 9 (11). United States: e1003326. doi:10.1371/journal.pcbi.1003326.
- Bailey, Timothy L. 2011. "DREME: Motif Discovery in Transcription Factor ChIP-Seq Data." *Bioinformatics (Oxford, England)* 27 (12). England: 1653–59. doi:10.1093/bioinformatics/btr261.
- Bailey, Timothy L, and Philip Machanick. 2012. "Inferring Direct DNA Binding from ChIP-Seq." *Nucleic Acids Research* 40 (17). England: e128. doi:10.1093/nar/gks433.
- Bailey, Timothy L, Nadya Williams, Chris Mischak, and Wilfred W Li. 2006. "MEME: Discovering and Analyzing DNA and Protein Sequence Motifs." *Nucleic Acids Research*. doi:10.1093/nar/gkl198.
- Bao, Hua, Hui Guo, Jinwei Wang, Renchao Zhou, Xuemei Lu, and Suhua Shi. 2009. "MapView: Visualization of Short Reads Alignment on a Desktop Computer." *Bioinformatics (Oxford, England)* 25 (12). England: 1554–55. doi:10.1093/bioinformatics/btp255.
- Bao, Suying, Rui Jiang, WingKeung Kwan, BinBin Wang, Xu Ma, and You-Qiang Song. 2011. "Evaluation of next-Generation Sequencing Software in Mapping and Assembly." *Journal of Human Genetics* 56 (6). England: 406–14. doi:10.1038/jhg.2011.43.
- Bardet, Anais F, Qiye He, Julia Zeitlinger, and Alexander Stark. 2012a. "A Computational Pipeline for Comparative ChIP-Seq Analyses." *Nature Protocols* 7 (1). England: 45–61. doi:10.1038/nprot.2011.420.
- Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. "High-Resolution Profiling of Histone Methylations in the Human Genome." *Cell* 129 (4). United States: 823–37. doi:10.1016/j.cell.2007.05.009.
- Bell, Oliver, Vijay K Tiwari, Nicolas H Thoma, and Dirk Schubeler. 2011. "Determinants and Dynamics of Genome Accessibility." *Nature Reviews. Genetics* 12 (8). England: 554–64. doi:10.1038/nrg3017.
- Benjamini, Yoav, and Yosef Hochberg. 1995a. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1). Wiley for the Royal Statistical Society: 289–300. doi:10.2307/2346101.
- Berger, Shelley L. 2002. "Histone Modifications in Transcriptional Regulation." *Current Opinion in Genetics & Development* 12 (2). England: 142–48.
- Berka, J, Y J Chen, J H Leamon, S Lefkowitz, K Lohman, V Makhijani, G J Sarkis, J Rothberg, M Weiner, and M Srinivasan. 2004. "Bead Emulsion Nucleic Acid Amplification." Google Patents. <https://www.google.com/patents/WO2004069849A2?cl=en>.
- Bernstein, Bradley E, Tarjei S Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J Huebert, James Cuff, Ben Fry, et al. 2006. "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells." *Cell* 125 (2). United States: 315–26. doi:10.1016/j.cell.2006.02.041.
- Bernstein, Bradley E, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, et al. 2010. "The NIH Roadmap Epigenomics Mapping Consortium." *Nature Biotechnology* 28 (10). United States: 1045–48. doi:10.1038/nbt1010-1045.
- Birney, Ewan, Jason D Lieb, Terrence S Furey, Gregory E Crawford, and Vishwanath R Iyer. 2010. "Allele-Specific and Heritable Chromatin Signatures in Humans." *Human Molecular Genetics* 19 (R2). England: R204–9. doi:10.1093/hmg/ddq404.

- Birney, Ewan, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigo, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, et al. 2007. "Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project." *Nature* 447 (7146). England: 799–816. doi:10.1038/nature05874.
- Blankenberg, Daniel, Assaf Gordon, Gregory Von Kuster, Nathan Coraor, James Taylor, and Anton Nekrutenko. 2010. "Manipulation of FASTQ Data with Galaxy." *Bioinformatics (Oxford, England)* 26 (14). England: 1783–85. doi:10.1093/bioinformatics/btq281.
- Blankenberg, Daniel, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. 2010. "Galaxy: A Web-Based Genome Analysis Tool for Experimentalists." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 19 (January). United States: Unit 19.10.1–21. doi:10.1002/0471142727.mb1910s89.
- Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics (Oxford, England)* 30 (15). England: 2114–20. doi:10.1093/bioinformatics/btu170.
- Bos, R, H Zhong, C F Hanrahan, E C Mommers, G L Semenza, H M Pinedo, M D Abeloff, J W Simons, P J van Diest, and E van der Wall. 2001. "Levels of Hypoxia-Inducible Factor-1 Alpha during Breast Carcinogenesis." *Journal of the National Cancer Institute* 93 (4). United States: 309–14.
- Boyle, Alan P, Justin Guinney, Gregory E Crawford, and Terrence S Furey. 2008. "F-Seq: A Feature Density Estimator for High-Throughput Sequence Tags." *Bioinformatics* 24 (21): 2537–38. doi:10.1093/bioinformatics/btn480 .
- Bradley, Robert K, Xiao-Yong Li, Cole Trapnell, Stuart Davidson, Lior Pachter, Hou Cheng Chu, Leath A Tonkin, Mark D Biggin, and Michael B Eisen. 2010a. "Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related *Drosophila* Species." *PLoS Biology* 8 (3). United States: e1000343. doi:10.1371/journal.pbio.1000343.
- Bryne, Jan Christian, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. 2008. "JASPAR, the Open Access Database of Transcription Factor-Binding Profiles: New Content and Tools in the 2008 Update." *Nucleic Acids Research* 36 (Database issue). England: D102–6. doi:10.1093/nar/gkm955.
- Bunone, G, P A Briand, R J Miksicek, and D Picard. 1996. "Activation of the Unliganded Estrogen Receptor by EGF Involves the MAP Kinase Pathway and Direct Phosphorylation." *The EMBO Journal*.
- Burkhardt, Stefan, Andreas Crauser, Paolo Ferragina, Hans-Peter Lenhof, Eric Rivals, and Martin Vingron. 1999. "Q-Gram Based Database Searching Using a Suffix Array (QUASAR)." In *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, 77–83. RECOMB '99. New York, NY, USA: ACM. doi:10.1145/299432.299460.
- Carroll, Jason S, X Shirley Liu, Alexander S Brodsky, Wei Li, Clifford A Meyer, Anna J Szary, Jerome Eeckhoutte, et al. 2005. "Chromosome-Wide Mapping of Estrogen Receptor Binding Reveals Long-Range Regulation Requiring the Forkhead Protein FoxA1." *Cell* 122 (1). United States: 33–43. doi:10.1016/j.cell.2005.05.008.

- Carroll, Jason S, Clifford A Meyer, Jun Song, Wei Li, Timothy R Geistlinger, Jerome Eeckhoutte, Alexander S Brodsky, et al. 2006. "Genome-Wide Analysis of Estrogen Receptor Binding Sites." *Nat Genet* 38 (11): 1289–97. <http://dx.doi.org/10.1038/ng1901>.
- Chedotal, Alain. 2007. "Chemotropic Axon Guidance Molecules in Tumorigenesis." *Progress in Experimental Tumor Research* 39. Switzerland: 78–90. doi:10.1159/0000100048.
- Chen, Xi, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B Vega, Eleanor Wong, et al. 2008. "Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells." *Cell* 133 (6). United States: 1106–17. doi:10.1016/j.cell.2008.04.043.
- Chen, Yangho, Tade Souzaiaia, and Ting Chen. 2009. "PerM: Efficient Mapping of Short Sequencing Reads with Periodic Full Sensitive Spaced Seeds." *Bioinformatics* 25 (19): 2514–21. doi:10.1093/bioinformatics/btp486.
- Chen, Yiwen, Clifford A Meyer, Tao Liu, Wei Li, Jun S Liu, and Xiaole Shirley Liu. 2011. "MM-ChIP Enables Integrative Analysis of Cross-Platform and between-Laboratory ChIP-Chip or ChIP-Seq Data." *Genome Biology* 12 (2). England: R11. doi:10.1186/gb-2011-12-2-r11.
- Chen, Yiwen, Nicolas Negre, Qunhua Li, Joanna O Mieczkowska, Matthew Slattery, Tao Liu, Yong Zhang, et al. 2012. "Systematic Evaluation of Factors Influencing ChIP-Seq Fidelity." *Nature Methods* 9 (6). United States: 609–14. doi:10.1038/nmeth.1985.
- Cheung, Ming-Sin, Thomas A Down, Isabel Latorre, and Julie Ahringer. 2011. "Systematic Bias in High-Throughput Sequencing Data and Its Correction by BEADS." *Nucleic Acids Research* 39 (15). England: e103. doi:10.1093/nar/gkr425.
- Chung, Dongjun, Pei Fen Kuan, Bo Li, Rajendran Sanalkumar, Kun Liang, Emery H Bresnick, Colin Dewey, and Sunduz Kees. 2011. "Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data." *PLoS Computational Biology* 7 (7). United States: e1002111. doi:10.1371/journal.pcbi.1002111.
- Chung, Dongjun, Dan Park, Kevin Myers, Jeffrey Grass, Patricia Kiley, Robert Landick, and Sunduz Kees. 2013. "dPeak: High Resolution Identification of Transcription Factor Binding Sites from PET and SET ChIP-Seq Data." *PLoS Computational Biology* 9 (10). United States: e1003246. doi:10.1371/journal.pcbi.1003246.
- Chung, Yih-Lin, Meei-Ling Sheu, Shun-Chun Yang, Chi-Hung Lin, and Sang-Hue Yen. 2002. "Resistance to Tamoxifen-Induced Apoptosis Is Associated with Direct Interaction between Her2/neu and Cell Membrane Estrogen Receptor in Breast Cancer." *International Journal of Cancer. Journal International Du Cancer* 97 (3). United States: 306–12.
- Cibulskis, Kristian, Aaron McKenna, Tim Fennell, Eric Banks, Mark DePristo, and Gad Getz. 2011. "ContEst: Estimating Cross-Contamination of Human Samples in next-Generation Sequencing Data." *Bioinformatics (Oxford, England)* 27 (18). England: 2601–2. doi:10.1093/bioinformatics/btr446.
- Ciruelos Gil, Eva Maria. 2014. "Targeting the PI3K/AKT/mTOR Pathway in Estrogen Receptor-Positive Breast Cancer." *Cancer Treatment Reviews* 40 (7): 862–71. doi:<http://dx.doi.org/10.1016/j.ctrv.2014.03.004>.
- Cock, Peter J A, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. 2010. "The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants." *Nucleic Acids Research*. doi:10.1093/nar/gkp1137.

- ConsortiumInternational, Human Genome Sequencing. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431 (7011). Macmillian Magazines Ltd.: 931–45. <http://dx.doi.org/10.1038/nature03001>.
- Couse, J F, and K S Korach. 1999. "Estrogen Receptor Null Mice: What Have We Learned and Where Will They Lead Us?" *Endocrine Reviews* 20 (3). UNITED STATES: 358–417. doi:10.1210/edrv.20.3.0370.
- Cox, Murray P, Daniel A Peterson, and Patrick J Biggs. 2010. "SolexaQA: At-a-Glance Quality Assessment of Illumina Second-Generation Sequencing Data." *BMC Bioinformatics* 11. England: 485. doi:10.1186/1471-2105-11-485.
- Creyghton, Menno P, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, et al. 2010. "Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State." *Proceedings of the National Academy of Sciences of the United States of America* 107 (50). United States: 21931–36. doi:10.1073/pnas.1016071107.
- Dahl, John Arne, and Philippe Collas. 2008. "MicroChIP--a Rapid Micro Chromatin Immunoprecipitation Assay for Small Cell Samples and Biopsies." *Nucleic Acids Research* 36 (3). England: e15. doi:10.1093/nar/gkm1158.
- Dai, Manhong, Robert C Thompson, Christopher Maher, Rafael Contreras-Galindo, Mark H Kaplan, David M Markovitz, Gil Omenn, and Fan Meng. 2010. "NGSQC: Cross-Platform Quality Analysis Pipeline for Deep Sequencing Data." *BMC Genomics* 11 Suppl 4. England: S7. doi:10.1186/1471-2164-11-S4-S7.
- Daley, Timothy, and Andrew D Smith. 2013. "Predicting the Molecular Complexity of Sequencing Libraries." *Nature Methods*. doi:10.1038/nmeth.2375.
- Dohm, Juliane C, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. 2008. "Substantial Biases in Ultra-Short Read Data Sets from High-Throughput DNA Sequencing." *Nucleic Acids Research*. doi:10.1093/nar/gkn425.
- Dolan, Peter C, and Dee R Denver. 2008a. "TileQC: A System for Tile-Based Quality Control of Solexa Data." *BMC Bioinformatics* 9. England: 250. doi:10.1186/1471-2105-9-250.
- Donlin, Maureen J. 2009. "Using the Generic Genome Browser (GBrowse)." *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxeavanis ... [et Al.]* Chapter 9 (December). United States: Unit 9.9. doi:10.1002/0471250953.bi0909s28.
- Eaton, Matthew L, Joseph A Prinz, Heather K MacAlpine, George Tretyakov, Peter V Kharchenko, and David M MacAlpine. 2011. "Chromatin Signatures of the Drosophila Replication Program." *Genome Research* 21 (2). United States: 164–74. doi:10.1101/gr.116038.110.
- Eckert, Lynn B, Gretchen A Repasky, Aylin S Ülkü, Aidan McFall, Hong Zhou, Carolyn I Sartor, and Channing J Der. 2004. "Involvement of Ras Activation in Human Breast Cancer Cell Signaling, Invasion, and Anoikis." *Cancer Research* 64 (13): 4585–92. doi:10.1158/0008-5472.CAN-04-0396 .
- Ernst, Jason, and Manolis Kellis. 2010. "Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome." *Nature Biotechnology* 28 (8). United States: 817–25. doi:10.1038/nbt.1662.

- Ernst, Jason, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shores, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, et al. 2011a. "Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types." *Nature* 473 (7345). England: 43–49. doi:10.1038/nature09906.
- Fedorova, Elena, and Daniele Zink. 2008. "Nuclear Architecture and Gene Regulation." *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1783 (11): 2174–84. doi:http://dx.doi.org/10.1016/j.bbamcr.2008.07.018.
- Fejes, Anthony P, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven J M Jones. 2008a. "FindPeaks 3.1: A Tool for Identifying Areas of Enrichment from Massively Parallel Short-Read Sequencing Technology." *Bioinformatics (Oxford, England)* 24 (15). England: 1729–30. doi:10.1093/bioinformatics/btn305.
- Feng, Dan, Tao Liu, Zheng Sun, Anne Bugge, Shannon E Mullican, Theresa Alenghat, X Shirley Liu, and Mitchell A Lazar. 2011. "A Circadian Rhythm Orchestrated by Histone Deacetylase 3 Controls Hepatic Lipid Metabolism." *Science (New York, N.Y.)* 331 (6022). United States: 1315–19. doi:10.1126/science.1198125.
- Feng, Xin, Robert Grossman, and Lincoln Stein. 2011. "PeakRanger: A Cloud-Enabled Peak Caller for ChIP-Seq Data." *BMC Bioinformatics* 12. England: 139. doi:10.1186/1471-2105-12-139.
- Ferragina, P, and G Manzini. 2000. "Opportunistic Data Structures with Applications." In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 390 – . FOCS '00. Washington, DC, USA: IEEE Computer Society. http://dl.acm.org/citation.cfm?id=795666.796543.
- Flicek, Paul. 2009. "The Need for Speed." *Genome Biology*. doi:10.1186/gb-2009-10-3-212.
- Flicek, Paul, and Ewan Birney. 2009. "Sense from Sequence Reads: Methods for Alignment and Assembly." *Nature Methods* 6 (11 Suppl). United States: S6–12. doi:10.1038/nmeth.1376.
- Foat, Barrett C, Alexandre V Morozov, and Harmen J Bussemaker. 2006. "Statistical Mechanical Modeling of Genome-Wide Transcription Factor Occupancy Data by MatrixREDUCE." *Bioinformatics* 22 (14): e141–49. http://bioinformatics.oxfordjournals.org/content/22/14/e141.abstract.
- Freedman, M, J San Martin, J O'Gorman, S Eckert, M E Lippman, S C Lo, E L Walls, and J Zeng. 2001. "Digitized Mammography: A Clinical Trial of Postmenopausal Women Randomly Assigned to Receive Raloxifene, Estrogen, or Placebo." *Journal of the National Cancer Institute* 93 (1). UNITED STATES: 51–56.
- Fried, M, and D M Crothers. 1981. "Equilibria and Kinetics of Lac Repressor-Operator Interactions by Polyacrylamide Gel Electrophoresis." *Nucleic Acids Research*.
- Froehlich, T, D Heindl, and A Roesler. 2010. "Miniaturized, High-Throughput Nucleic Acid Analysis." Google Patents. https://www.google.com/patents/US20100248237.
- Fujita, Pauline A, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, et al. 2011. "The UCSC Genome Browser Database: Update 2011." *Nucleic Acids Research*. doi:10.1093/nar/gkq963.
- Fujiwara, Tohru, Henriette O'Geen, Sunduz Keles, Kimberly Blahnik, Amelia K Linnemann, Yoon-A Kang, Kyunghye Choi, Peggy J Farnham, and Emery H Bresnick. 2009. "Discovering Hematopoietic Mechanisms through Genome-Wide Analysis of GATA Factor Chromatin Occupancy." *Molecular Cell* 36 (4). United States: 667–81. doi:10.1016/j.molcel.2009.11.001.

- Furey, Terrence S. 2012. "ChIP-Seq and beyond: New and Improved Methodologies to Detect and Characterize Protein-DNA Interactions." *Nature Reviews. Genetics* 13 (12). England: 840–52. doi:10.1038/nrg3306.
- Galas, D J, and A Schmitz. 1978. "DNase Footprinting: A Simple Method for the Detection of Protein-DNA Binding Specificity." *Nucleic Acids Research* 5 (9). ENGLAND: 3157–70.
- Galinsky, V L. 2012. "YOABS: Yet Other Aligner of Biological Sequences—an Efficient Linearly Scaling Nucleotide Aligner." *Bioinformatics* 28 (8): 1070–77. doi:10.1093/bioinformatics/bts102
- Garber, Manuel, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. 2011. "Computational Methods for Transcriptome Annotation and Quantification Using RNA-Seq." *Nature Methods* 8 (6). United States: 469–77. doi:10.1038/nmeth.1613.
- Garner, M M, and A Revzin. 1981. "A Gel Electrophoresis Method for Quantifying the Binding of Proteins to Specific DNA Regions: Application to Components of the Escherichia Coli Lactose Operon Regulatory System." *Nucleic Acids Research*.
- Gerstein, Mark B, Zhi John Lu, Eric L Van Nostrand, Chao Cheng, Bradley I Arshinoff, Tao Liu, Kevin Y Yip, et al. 2010. "Integrative Analysis of the Caenorhabditis Elegans Genome by the modENCODE Project." *Science (New York, N.Y.)* 330 (6012). United States: 1775–87. doi:10.1126/science.1196914.
- Gerstein, Mark B, Can Bruce, Joel S Rozowsky, Deyou Zheng, Jiang Du, Jan O Korbel, Olof Emanuelsson, Zhengdong D Zhang, Sherman Weissman, and Michael Snyder. 2007. "What Is a Gene, Post-ENCODE? History and Updated Definition." *Genome Research* 17 (6): 669–81. doi:10.1101/gr.6339607 .
- Geyer, P K, and V G Corces. 1992. "DNA Position-Specific Repression of Transcription by a Drosophila Zinc Finger Protein." *Genes & Development* 6 (10). UNITED STATES: 1865–73.
- Giardine, Belinda, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, et al. 2005. "Galaxy: A Platform for Interactive Large-Scale Genome Analysis." *Genome Research*. doi:10.1101/gr.4086505.
- Goecks, Jeremy, Anton Nekrutenko, and James Taylor. 2010. "Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences." *Genome Biology* 11 (8). England: R86. doi:10.1186/gb-2010-11-8-r86.
- Gräf, Stefan, Fiona G G Nielsen, Stefan Kurtz, Martijn A Huynen, Ewan Birney, Henk Stunnenberg, and Paul Flicek. 2007. "Optimized Design and Assessment of Whole Genome Tiling Arrays." *Bioinformatics* 23 (13): i195–204. doi:10.1093/bioinformatics/btm200 .
- Grant, Charles E, Timothy L Bailey, and William Stafford Noble. 2011. "FIMO: Scanning for Occurrences of a given Motif." *Bioinformatics (Oxford, England)* 27 (7). England: 1017–18. doi:10.1093/bioinformatics/btr064.
- Green, S, P Walter, V Kumar, A Krust, J M Bornert, P Argos, and P Chambon. 1986. "Human Oestrogen Receptor cDNA: Sequence, Expression and Homology to v-Erb-A." *Nature* 320 (6058). ENGLAND: 134–39. doi:10.1038/320134a0.

- Gregoret, Ivan V. n.d. "No Title."
- Gronemeyer, H. 1991. "Transcription Activation by Estrogen and Progesterone Receptors." *Annual Review of Genetics* 25. UNITED STATES: 89–123. doi:10.1146/annurev.ge.25.120191.000513.
- Gronemeyer, H, and V Laudet. 1995. "Transcription Factors 3: Nuclear Receptors." *Protein Profile* 2 (11). ENGLAND: 1173–1308.
- Guo, Gao, Sebastian Bauer, Jochen Hecht, Marcel H Schulz, Andreas Busche, and Peter N Robinson. 2008. "A Short Ultraconserved Sequence Drives Transcription from an Alternate FBN1 Promoter." *The International Journal of Biochemistry & Cell Biology* 40 (4). England: 638–50. doi:10.1016/j.biocel.2007.09.004.
- Guo, Yuchun, Georgios Papachristoudis, Robert C Altshuler, Georg K Gerber, Tommi S Jaakkola, David K Gifford, and Shaun Mahony. 2010. "Discovering Homotypic Binding Events at High Spatial Resolution." *Bioinformatics (Oxford, England)* 26 (24). England: 3028–34. doi:10.1093/bioinformatics/btq590.
- Gupta, Shobhit, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. 2007. "Quantifying Similarity between Motifs." *Genome Biology* 8 (2). England: R24. doi:10.1186/gb-2007-8-2-r24.
- Gyorffy, Balazs, Zsombor Benke, Andras Lanczky, Balint Balazs, Zoltan Szallasi, Jozsef Timar, and Reinhold Schafer. 2012. "RecurrenceOnline: An Online Analysis Tool to Determine Breast Cancer Recurrence and Hormone Receptor Status Using Microarray Data." *Breast Cancer Research and Treatment* 132 (3). Netherlands: 1025–34. doi:10.1007/s10549-011-1676-y.
- Hach, Faraz, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E Eichler, and S Cenik Sahinalp. 2010. "mrsFAST: A Cache-Oblivious Algorithm for Short-Read Mapping." *Nature Methods*. United States. doi:10.1038/nmeth0810-576.
- Harvey, Kieran F, Xiaomeng Zhang, and David M Thomas. 2013. "The Hippo Pathway and Human Cancer." *Nature Reviews. Cancer* 13 (4). England: 246–57. doi:10.1038/nrc3458.
- He, Bing, Changya Chen, Li Teng, and Kai Tan. 2014. "Global View of Enhancer–promoter Interactome in Human Cells." *Proceedings of the National Academy of Sciences of the United States of America* 111 (21). National Academy of Sciences: E2191–99. doi:10.1073/pnas.1320308111.
- He, Housheng Hansen, Clifford A Meyer, Hyunjin Shin, Shannon T Bailey, Gang Wei, Qianben Wang, Yong Zhang, et al. 2010a. "Nucleosome Dynamics Define Transcriptional Enhancers." *Nature Genetics* 42 (4). United States: 343–47. doi:10.1038/ng.545.
- Hecht, J, V Seitz, M Urban, F Wagner, P N Robinson, A Stiege, C Dieterich, et al. 2007. "Detection of Novel Skeletogenesis Target Genes by Comprehensive Analysis of a Runx2(–/–) Mouse Model." *Gene Expression Patterns : GEP* 7 (1-2). Netherlands: 102–12. doi:10.1016/j.modgep.2006.05.014.
- Heintzman, Nathaniel D, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, et al. 2007. "Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome." *Nature Genetics* 39 (3). United States: 311–18. doi:10.1038/ng1966.
- Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. 2010. "Simple Combinations

- of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4). United States: 576–89. doi:10.1016/j.molcel.2010.05.004.
- Heldring, Nina, Ashley Pike, Sandra Andersson, Jason Matthews, Guojun Cheng, Johan Hartman, Michel Tujague, et al. 2007. "Estrogen Receptors: How Do They Signal and What Are Their Targets." *Physiological Reviews* 87 (3). United States: 905–31. doi:10.1152/physrev.00026.2006.
- Hembruff, Stacey L, and Nikki Cheng. 2009. "Chemokine Signaling in Cancer: Implications on the Tumor Microenvironment and Therapeutic Targeting." *Cancer Therapy* 7 (A): 254–67.
- Henttu, P M, E Kalkhoven, and M G Parker. 1997. "AF-2 Activity and Recruitment of Steroid Receptor Coactivator 1 to the Estrogen Receptor Depend on a Lysine Residue Conserved in Nuclear Receptors." *Molecular and Cellular Biology* 17 (4). UNITED STATES: 1832–39.
- Hertz, G Z, and G D Stormo. 1999. "Identifying DNA and Protein Patterns with Statistically Significant Alignments of Multiple Sequences." *Bioinformatics (Oxford, England)* 15 (7-8). ENGLAND: 563–77.
- Herynk, Matthew H, and Suzanne A W Fuqua. 2004. "Estrogen Receptor Mutations in Human Disease." *Endocrine Reviews* 25 (6). United States: 869–98. doi:10.1210/er.2003-0010.
- Hewitt, Sylvia Curtis, and Kenneth S Korach. 2002. "Estrogen Receptors: Structure, Mechanisms and Function." *Reviews in Endocrine & Metabolic Disorders* 3 (3). United States: 193–200.
- Hillier, LaDeana W, Gabor T Marth, Aaron R Quinlan, David Dooling, Ginger Fewell, Derek Barnett, Paul Fox, et al. 2008. "Whole-Genome Sequencing and Variant Discovery in *C. Elegans*." *Nature Methods* 5 (2). United States: 183–88. doi:10.1038/nmeth.1179.
- Hoffman, Michael M, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. 2012. "Unsupervised Pattern Discovery in Human Chromatin Structure through Genomic Segmentation." *Nature Methods* 9 (5). United States: 473–76. doi:10.1038/nmeth.1937.
- Hoffmann, Steve, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jorg Vogel, Peter F Stadler, and Jorg Hackermuller. 2009. "Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures." *PLoS Computational Biology* 5 (9). United States: e1000502. doi:10.1371/journal.pcbi.1000502.
- Homer, Nils, Barry Merriman, and Stanley F Nelson. 2009. "BFAST: An Alignment Tool for Large Scale Genome Resequencing." *PloS One* 4 (11). United States: e7767. doi:10.1371/journal.pone.0007767.
- Huang, Da Wei, Brad T Sherman, and Richard A Lempicki. 2009a. "Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources." *Nature Protocols* 4 (1). England: 44–57. doi:10.1038/nprot.2008.211.
- Huang, Weichun, and Gabor Marth. 2008. "EagleView: A Genome Assembly Viewer for next-Generation Sequencing Technologies." *Genome Research* 18 (9). United States: 1538–43. doi:10.1101/gr.076067.108.

- Hurtado, Antoni, Kelly A Holmes, Caryn S Ross-Innes, Dominic Schmidt, and Jason S Carroll. 2011. "FOXA1 Is a Key Determinant of Estrogen Receptor Function and Endocrine Response." *Nature Genetics* 43 (1). United States: 27–33. doi:10.1038/ng.730.
- Huse, Susan M, Julie A Huber, Hilary G Morrison, Mitchell L Sogin, and David Mark Welch. 2007. "Accuracy and Quality of Massively Parallel DNA Pyrosequencing." *Genome Biology* 8 (7). England: R143. doi:10.1186/gb-2007-8-7-r143.
- Itoh, Masahiko, Celeste M Nelson, Connie A Myers, and Mina J Bissell. 2007. "Rap1 Integrates Tissue Polarity, Lumen Formation, and Tumorigenic Potential in Human Breast Epithelial Cells." *Cancer Research* 67 (10): 4759–66. doi:10.1158/0008-5472.CAN-06-4246 .
- Iyer, V R, C E Horak, C S Scafe, D Botstein, M Snyder, and P O Brown. 2001. "Genomic Binding Sites of the Yeast Cell-Cycle Transcription Factors SBF and MBF." *Nature* 409 (6819). England: 533–38. doi:10.1038/35054095.
- Jackson, V. 1978. "Studies on Histone Organization in the Nucleosome Using Formaldehyde as a Reversible Cross-Linking Agent." *Cell* 15 (3). ENGLAND: 945–54.
- Jeong, Jaesik, Lang Li, Yunlong Liu, Kenneth Nephew, Tim Huang, and Changyu Shen. 2010. "An Empirical Bayes Model for Gene Expression and Methylation Profiles in Antiestrogen Resistant Breast Cancer." *BMC Medical Genomics* 3 (1): 55. <http://www.biomedcentral.com/1755-8794/3/55>.
- Ji, Hongkai. 2010. "Computational Analysis of ChIP-Seq Data." *Methods in Molecular Biology (Clifton, N.J.)* 674. United States: 143–59. doi:10.1007/978-1-60761-854-6_9.
- Ji, Hongkai, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. 2008a. "An Integrated Software System for Analyzing ChIP-Chip and ChIP-Seq Data." *Nature Biotechnology* 26 (11). United States: 1293–1300. doi:10.1038/nbt.1505.
- Ji, Hongkai, Hui Jiang, Wenxiu Ma, and Wing Hung Wong. 2011. "Using CisGenome to Analyze ChIP-Chip and ChIP-Seq Data." *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxeavanis ... [et Al.]* Chapter 2 (March). United States: Unit2.13. doi:10.1002/0471250953.bi0213s33.
- Ji, Xuwo, Wei Li, Jun Song, Liping Wei, and X Shirley Liu. 2006. "CEAS: Cis-Regulatory Element Annotation System." *Nucleic Acids Research* 34 (suppl 2): W551–54. doi:10.1093/nar/gkl322 .
- Jiang, Hui, and Wing Hung Wong. 2008. "SeqMap: Mapping Massive Amount of Oligonucleotides to the Genome." *Bioinformatics (Oxford, England)* 24 (20). England: 2395–96. doi:10.1093/bioinformatics/btn429.
- Jiang, Peng, Hongfang Wang, Wei Li, Chongzhi Zang, Bo Li, Yinling J Wong, Cliff Meyer, Jun S Liu, Jon C Aster, and X Shirley Liu. 2015. "Network Analysis of Gene Essentiality in Functional Genomics Experiments." *Genome Biology* 16 (October). London: BioMed Central: 239. doi:10.1186/s13059-015-0808-9.
- Jiang, Peng, Matthew L Freedman, Jun S Liu, and Xiaole Shirley Liu. 2015. "Inference of Transcriptional Regulation in Cancers." *Proceedings of the National Academy of Sciences of the United States of America* 112 (25). National Academy of Sciences: 7731–36. doi:10.1073/pnas.1424272112.

- Johnson, David S, Ali Mortazavi, Richard M Myers, and Barbara Wold. 2007. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." *Science* 316 (5830): 1497–1502. doi:10.1126/science.1141319 .
- Jothi, Raja, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. 2008a. "Genome-Wide Identification of in Vivo Protein-DNA Binding Sites from ChIP-Seq Data." *Nucleic Acids Research* 36 (16). England: 5221–31. doi:10.1093/nar/gkn488.
- Kahlert, S, S Nuedling, M van Eickels, H Vetter, R Meyer, and C Grohe. 2000. "Estrogen Receptor Alpha Rapidly Activates the IGF-1 Receptor Pathway." *The Journal of Biological Chemistry* 275 (24). UNITED STATES: 18447–53. doi:10.1074/jbc.M910345199.
- Kallin, Eric M, Ru Cao, Raja Jothi, Kai Xia, Kairong Cui, Keji Zhao, and Yi Zhang. 2009. "Genome-Wide uH2A Localization Analysis Highlights Bmi1-Dependent Deposition of the Mark at Repressed Genes." Edited by Dirk Schübeler. *PLoS Genetics*. San Francisco, USA. doi:10.1371/journal.pgen.1000506.
- Kärkkäinen, Juha. 2007. "Fast BWT in Small Space by Blockwise Suffix Sorting." *Theor. Comput. Sci.* 387 (3). Essex, UK: Elsevier Science Publishers Ltd.: 249–57. doi:10.1016/j.tcs.2007.07.018.
- Karolchik, Donna, Angela S Hinrichs, Terrence S Furey, Krishna M Roskin, Charles W Sugnet, David Haussler, and W James Kent. 2004. "The UCSC Table Browser Data Retrieval Tool." *Nucleic Acids Research* 32 (Database issue). England: D493–96. doi:10.1093/nar/gkh103.
- Kasowski, Maya, Fabian Grubert, Christopher Heffelfinger, Manoj Hariharan, Akwasi Asabere, Sebastian M Waszak, Lukas Habegger, et al. 2010. "Variation in Transcription Factor Binding among Humans." *Science (New York, N.Y.)* 328 (5975). United States: 232–35. doi:10.1126/science.1183621.
- Kel, A E, E Gossling, I Reuter, E Cheremushkin, O V Kel-Margoulis, and E Wingender. 2003. "MATCH: A Tool for Searching Transcription Factor Binding Sites in DNA Sequences." *Nucleic Acids Research* 31 (13). England: 3576–79.
- Kellum, R, and P Schedl. 1992. "A Group of Scs Elements Function as Domain Boundaries in an Enhancer-Blocking Assay." *Molecular and Cellular Biology* 12 (5): 2424–31.
- Kent, W James. 2002. "BLAT--the BLAST-like Alignment Tool." *Genome Research* 12 (4). United States: 656–64.
- Kent, W James, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. 2002. "The Human Genome Browser at UCSC." *Genome Research* 12 (6). United States: 996–1006.
- Kharchenko, Peter V, Michael Y Tolstorukov, and Peter J Park. 2008a. "Design and Analysis of ChIP-Seq Experiments for DNA-Binding Proteins." *Nature Biotechnology* 26 (12). United States: 1351–59. doi:10.1038/nbt.1508.
- Kim, Tae Hoon, Leah O Barrera, Ming Zheng, Chunxu Qu, Michael A Singer, Todd A Richmond, Yingnian Wu, Roland D Green, and Bing Ren. 2005. "A High-Resolution Map of Active Promoters in the Human Genome." *Nature* 436 (7052). England: 876–80. doi:10.1038/nature03877.
- Kim, Tae Hoon, and Bing Ren. 2006. "Genome-Wide Analysis of Protein-DNA Interactions." *Annual Review of Genomics and Human Genetics* 7. United States: 81–102. doi:10.1146/annurev.genom.7.080505.115634.

- Kinsinger, Linda S, Russell Harris, Steven H Woolf, Harold C Sox, and Kathleen N Lohr. 2002. "Chemoprevention of Breast Cancer: A Summary of the Evidence for the U.S. Preventive Services Task Force." *Annals of Internal Medicine* 137 (1). United States: 59–69.
- Kircher, Martin, Patricia Heyn, and Janet Kelso. 2011. "Addressing Challenges in the Production and Analysis of Illumina Sequencing Data." *BMC Genomics* 12. England: 382. doi:10.1186/1471-2164-12-382.
- Klus, Petr, Simon Lam, Dag Lyberg, MingSin Cheung, Graham Pullan, Ian McFarlane, GilesSH Yeo, and BrianYH Lam. 2012. "BarraCUDA - a Fast Short Read Sequence Aligner Using Graphics Processing Units." *BMC Research Notes* 5 (1). BioMed Central: 1–7. doi:10.1186/1756-0500-5-27.
- Knowlden, Janice M, Iain R Hutcheson, Helen E Jones, Tracieann Madden, Julia M W Gee, Maureen E Harper, Denise Barrow, Alan E Wakeling, and Robert I Nicholson. 2003. "Elevated Levels of Epidermal Growth Factor Receptor/c-erbB2 Heterodimers Mediate an Autocrine Growth Regulatory Pathway in Tamoxifen-Resistant MCF-7 Cells." *Endocrinology* 144 (3). United States: 1032–44. doi:10.1210/en.2002-220620.
- Kolasinska-Zwierz, Paulina, Thomas Down, Isabel Latorre, Tao Liu, X Shirley Liu, and Julie Ahringer. 2009. "Differential Chromatin Marking of Introns and Expressed Exons by H3K36me3." *Nature Genetics* 41 (3). United States: 376–81. doi:10.1038/ng.322.
- Korbel, Jan O, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, et al. 2007. "Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome." *Science (New York, N.Y.)* 318 (5849). United States: 420–26. doi:10.1126/science.1149504.
- Krawitz, Peter, Christian Rodelsperger, Marten Jager, Luke Jostins, Sebastian Bauer, and Peter N Robinson. 2010. "Microindel Detection in Short-Read Sequence Data." *Bioinformatics (Oxford, England)* 26 (6). England: 722–29. doi:10.1093/bioinformatics/btq027.
- Krust, A, S Green, P Argos, V Kumar, P Walter, J M Bornert, and P Chambon. 1986. "The Chicken Oestrogen Receptor Sequence: Homology with v-erbA and the Human Oestrogen and Glucocorticoid Receptors." *The EMBO Journal* 5 (5): 891–97. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1166879/>.
- Kucherov, Gregory, Laurent Noé, and Mikhail Roytberg. 2006. "A Unifying Framework for Seed Sensitivity and Its Application to Subset Seeds." *Journal of Bioinformatics and Computational Biology*. doi:10.1142/S0219720006001977.
- Kulakovskiy, Ivan V, Ilya E Vorontsov, Ivan S Yevshin, Anastasiia V Soboleva, Artem S Kasianov, Haitham Ashoor, Wail Ba-alawi, et al. 2016. "HOCOMOCO: Expansion and Enhancement of the Collection of Transcription Factor Binding Sites Models." *Nucleic Acids Research* 44 (Database issue). Oxford University Press: D116–25. doi:10.1093/nar/gkv1249.
- Kulakovskiy, I V, V A Boeva, A V Favorov, and V J Makeev. 2010. "Deep and Wide Digging for Binding Motifs in ChIP-Seq Data." *Bioinformatics (Oxford, England)* 26 (20). England: 2622–23. doi:10.1093/bioinformatics/btq488.
- Kumar, V, and P Chambon. 1988. "The Estrogen Receptor Binds Tightly to Its Responsive Element as a Ligand-Induced Homodimer." *Cell* 55 (1). UNITED STATES: 145–56.
- Kumar, V, S Green, G Stack, M Berry, J R Jin, and P Chambon. 1987. "Functional Domains of the Human Estrogen Receptor." *Cell* 51 (6). UNITED STATES: 941–51.

- Kuttippurathu, Lakshmi, Michael Hsing, Yongchao Liu, Bertil Schmidt, Douglas L Maskell, Kyungjoon Lee, Aibin He, William T Pu, and Sek Won Kong. 2011. "CompleteMOTIFs: DNA Motif Discovery Platform for Transcription Factor Binding Experiments." *Bioinformatics (Oxford, England)* 27 (5). England: 715–17. doi:10.1093/bioinformatics/btq707.
- Łabaj, Paweł P, Germán G Leparo, Bryan E Linggi, Lye Meng Markillie, H Steven Wiley, and David P Kreil. 2011. "Characterization and Improvement of RNA-Seq Precision in Quantitative Transcript Expression Profiling." *Bioinformatics* 27 (13): i383–91. <http://bioinformatics.oxfordjournals.org/content/27/13/i383.abstract>.
- Laganière, Josée, Geneviève Deblois, Céline Lefebvre, Alain R Bataille, François Robert, and Vincent Giguère. 2005. "Location Analysis of Estrogen Receptor α Target Promoters Reveals That FOXA1 Defines a Domain of the Estrogen Response." *Proceedings of the National Academy of Sciences of the United States of America* 102 (33). National Academy of Sciences: 11651–56. doi:10.1073/pnas.0505575102.
- Lamy, Philippe, Carsten Wiuf, Torben F Orntoft, and Claus L Andersen. 2011. "Rseg--an R Package to Optimize Segmentation of SNP Array Data." *Bioinformatics (Oxford, England)* 27 (3). England: 419–20. doi:10.1093/bioinformatics/btq668.
- Landt, Stephen G, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, et al. 2012. "ChIP-Seq Guidelines and Practices of the ENCODE and modENCODE Consortia." *Genome Research* 22 (9). United States: 1813–31. doi:10.1101/gr.136184.111.
- Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3). England: R25. doi:10.1186/gb-2009-10-3-r25.
- Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4). United States: 357–59. doi:10.1038/nmeth.1923.
- Le Romancer, Muriel, Coralie Poulard, Pascale Cohen, Stephanie Sentis, Jack-Michel Renoir, and Laura Corbo. 2011. "Cracking the Estrogen Receptor's Posttranslational Code in Breast Tumors." *Endocrine Reviews* 32 (5). United States: 597–622. doi:10.1210/er.2010-0016.
- Leblanc, B, and T Moss. 1994. "DNase I Footprinting." *Methods in Molecular Biology (Clifton, N.J.)* 30. UNITED STATES: 1–10. doi:10.1385/0-89603-256-6:1.
- Leclercq, Guy, Marc Lacroix, Ioanna Laios, and Guy Laurent. 2006. "Estrogen Receptor Alpha: Impact of Ligands on Intracellular Shuttling and Turnover Rate in Breast Cancer Cells." *Current Cancer Drug Targets* 6 (1). Netherlands: 39–64.
- Lefrançois, Philippe, Ghia M Euskirchen, Raymond K Auerbach, Joel Rozowsky, Theodore Gibson, Christopher M Yellman, Mark Gerstein, and Michael Snyder. 2009. "Efficient Yeast ChIP-Seq Using Multiplex Short-Read DNA Sequencing." *BMC Genomics* 10. England: 37. doi:10.1186/1471-2164-10-37.
- Lewis, S E, S M J Searle, N Harris, M Gibson, V Lyer, J Richter, C Wiel, et al. 2002a. "Apollo: A Sequence Annotation Editor." *Genome Biology* 3 (12). England: RESEARCH0082.

- Li, Guoliang, Melissa J Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, et al. 2010. "ChIA-PET Tool for Comprehensive Chromatin Interaction Analysis with Paired-End Tag Sequencing." *Genome Biology* 11 (2). England: R22. doi:10.1186/gb-2010-11-2-r22.
- Li, Heng, and Richard Durbin. 2009a. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics*. doi:10.1093/bioinformatics/btp324.
- . 2010. "Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform." *Bioinformatics (Oxford, England)* 26 (5). England: 589–95. doi:10.1093/bioinformatics/btp698.
- Li, Heng, and Nils Homer. 2010. "A Survey of Sequence Alignment Algorithms for next-Generation Sequencing." *Briefings in Bioinformatics* 11 (5): 473–83. doi:10.1093/bib/bbq015 .
- Li, Heng, Jue Ruan, and Richard Durbin. 2008. "Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores." *Genome Research*. doi:10.1101/gr.078212.108.
- Li, Leping. 2009. "Gadem: A Genetic Algorithm Guided Formation of Spaced Dyads Coupled with an EM Algorithm for Motif Discovery." *Journal of Computational Biology*. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA. doi:10.1089/cmb.2008.16TT.
- Li, Ming, Bin Ma, Derek Kisman, and John Tromp. 2003. "PatternHunter II: Highly Sensitive and Fast Homology Search." *Genome Informatics. International Conference on Genome Informatics* 14. Japan: 164–75.
- Li, Qunhua, James B Brown, Haiyan Huang, and Peter J Bickel. 2011. "Measuring Reproducibility of High-Throughput Experiments." *The Institute of Mathematical Statistics*, 1752–79. doi:10.1214/11-AOAS466.
- Li, Ruiqiang, Yingrui Li, Karsten Kristiansen, and Jun Wang. 2008. "SOAP: Short Oligonucleotide Alignment Program." *Bioinformatics* 24 (5): 713–14. doi:10.1093/bioinformatics/btn025 .
- Li, Ruiqiang, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. 2009a. "SOAP2: An Improved Ultrafast Tool for Short Read Alignment." *Bioinformatics (Oxford, England)* 25 (15). England: 1966–67. doi:10.1093/bioinformatics/btp336.
- Liang, Kun, and Sündüz Keleş. 2012. "Detecting Differential Binding of Transcription Factors with ChIP-Seq." *Bioinformatics* 28 (1): 121–22. <http://bioinformatics.oxfordjournals.org/content/28/1/121.abstract>.
- Liu, Chi-Man, Thomas Wong, Edward Wu, Ruibang Luo, Siu-Ming Yiu, Yingrui Li, Bingqiang Wang, et al. 2012. "SOAP3: Ultra-Fast GPU-Based Parallel Alignment Tool for Short Reads." *Bioinformatics* 28 (6): 878–79. doi:10.1093/bioinformatics/bts061 .
- Lin, Hao, Zefeng Zhang, Michael Q Zhang, Bin Ma, and Ming Li. 2008. "ZOOM! Zillions of Oligos Mapped." *Bioinformatics (Oxford, England)* 24 (21). England: 2431–37. doi:10.1093/bioinformatics/btn416.
- Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. 2012. "Comparison of next-Generation Sequencing Systems." *Journal of Biomedicine & Biotechnology* 2012. United States: 251364. doi:10.1155/2012/251364.
- Liu, Tao, Jorge A Ortiz, Len Taing, Clifford A Meyer, Bernett Lee, Yong Zhang, Hyunjin Shin, et al. 2011. "Cistrome: An Integrative Platform for Transcriptional Regulation Studies." *Genome Biology* 12 (8). England: R83. doi:10.1186/gb-2011-12-8-r83.

- Liu, Tao, Andreas Rechtsteiner, Thea A Egelhofer, Anne Vielle, Isabel Latorre, Ming-Sin Cheung, Sevinc Ercan, et al. 2011. "Broad Chromosomal Domains of Histone Modification Patterns in *C. Elegans*." *Genome Research* 21 (2). United States: 227–36. doi:10.1101/gr.115519.110.
- Liu, Wen, Bogdan Tanasa, Oksana V Tyurina, Tian Yuan Zhou, Reto Gassmann, Wei Ting Liu, Kenneth A Ohgi, et al. 2010. "PHF8 Mediates Histone H4 Lysine 20 Demethylation Events Involved in Cell Cycle Progression." *Nature* 466 (7305). England: 508–12. doi:10.1038/nature09272.
- Lun, Desmond S, Ashley Sherrid, Brian Weiner, David R Sherman, and James E Galagan. 2009. "A Blind Deconvolution Approach to High-Resolution Mapping of Transcription Factor Binding Sites from ChIP-Seq Data." *Genome Biology* 10 (12). England: R142. doi:10.1186/gb-2009-10-12-r142.
- Lunter, Gerton, and Martin Goodson. 2011. "Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads." *Genome Research* 21 (6). United States: 936–39. doi:10.1101/gr.111120.110.
- Lupien, Mathieu, Jerome Eeckhoutte, Clifford A Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S Carroll, X Shirley Liu, and Myles Brown. 2008. "FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription." *Cell* 132 (6). United States: 958–70. doi:10.1016/j.cell.2008.01.018.
- Lupien, Mathieu, Clifford A Meyer, Shannon T Bailey, Jerome Eeckhoutte, Jennifer Cook, Thomas Westerling, Xiaoyang Zhang, et al. 2010. "Growth Factor Stimulation Induces a Distinct ER(alpha) Cistrome Underlying Breast Cancer Endocrine Resistance." *Genes & Development* 24 (19). United States: 2219–27. doi:10.1101/gad.1944810.
- Ma, Bin, and Ming Li. 2007. "On the Complexity of the Spaced Seeds." *Journal of Computer and System Sciences* 73 (7): 1024–34. doi:http://dx.doi.org/10.1016/j.jcss.2007.03.008.
- Ma, Bin, John Tromp, and Ming Li. 2002. "PatternHunter: Faster and More Sensitive Homology Search." *Bioinformatics (Oxford, England)* 18 (3). England: 440–45.
- Ma, Xiaotu, Ashwinikumar Kulkarni, Zhihua Zhang, Zhenyu Xuan, Robert Serfling, and Michael Q Zhang. 2012. "A Highly Efficient and Effective Motif Discovery Method for ChIP-seq/ChIP-Chip Data Using Positional Information." *Nucleic Acids Research* 40 (7). England: e50. doi:10.1093/nar/gkr1135.
- Machanick, Philip, and Timothy L Bailey. 2011. "MEME-ChIP: Motif Analysis of Large DNA Datasets." *Bioinformatics (Oxford, England)* 27 (12). England: 1696–97. doi:10.1093/bioinformatics/btr189.
- Mader, S, P Chambon, and J H White. 1993. "Defining a Minimal Estrogen Receptor DNA Binding Domain." *Nucleic Acids Research*.
- Maher, Christopher A, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M Chinnaiyan. 2009. "Transcriptome Sequencing to Detect Gene Fusions in Cancer." *Nature* 458 (7234). England: 97–101. doi:10.1038/nature07638.
- Mahony, Shaun, and Panayiotis V Benos. 2007. "STAMP: A Web Tool for Exploring DNA-Binding Motif Similarities." *Nucleic Acids Research* 35 (Web Server issue). England: W253–58. doi:10.1093/nar/gkm272.

- Mardis, Elaine R. 2007. "ChIP-Seq: Welcome to the New Frontier." *Nature Methods*. United States. doi:10.1038/nmeth0807-613.
- . 2008a. "Next-Generation DNA Sequencing Methods." *Annual Review of Genomics and Human Genetics* 9. United States: 387–402. doi:10.1146/annurev.genom.9.081307.164359.
- . 2008b. "The Impact of next-Generation Sequencing Technology on Genetics." *Trends in Genetics : TIG* 24 (3). England: 133–41. doi:10.1016/j.tig.2007.12.007.
- Margulies, Marcel, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, et al. 2005. "Genome Sequencing in Microfabricated High-Density Picolitre Reactors." *Nature* 437 (7057). England: 376–80. doi:10.1038/nature03959.
- Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. 2008. "RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays." *Genome Research* 18 (9). United States: 1509–17. doi:10.1101/gr.079558.108.
- Martinez-Alcantara, A, E Ballesteros, C Feng, M Rojas, H Koshinsky, V Y Fofanov, P Havlak, and Y Fofanov. 2009. "PIQA: Pipeline for Illumina G1 Genome Analyzer Data Quality Assessment." *Bioinformatics (Oxford, England)* 25 (18). England: 2438–39. doi:10.1093/bioinformatics/btp429.
- Mateos, Julieta L, Pedro Madrigal, Kenichi Tsuda, Vimal Rawat, René Richter, Maida Romera-Branchat, Fabio Fornara, Korbinian Schneeberger, Pawel Krajewski, and George Coupland. 2015. "Combinatorial Activities of SHORT VEGETATIVE PHASE and FLOWERING LOCUS C Define Distinct Modes of Flowering Regulation in Arabidopsis." *Genome Biology*. London. doi:10.1186/s13059-015-0597-1.
- Matys, V, E Fricke, R Geffers, E Gossling, M Haubrock, R Hehl, K Hornischer, et al. 2003. "TRANSFAC: Transcriptional Regulation, from Patterns to Profiles." *Nucleic Acids Research* 31 (1). England: 374–78.
- McInerney, E M, K E Weis, J Sun, S Mosselman, and B S Katzenellenbogen. 1998. "Transcription Activation by the Human Estrogen Receptor Subtype Beta (ER Beta) Studied with ER Beta and ER Alpha Receptor Chimeras." *Endocrinology* 139 (11). UNITED STATES: 4513–22. doi:10.1210/endo.139.11.6298.
- McLean, Cory Y, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. 2010. "GREAT Improves Functional Interpretation of Cis-Regulatory Regions." *Nature Biotechnology* 28 (5). United States: 495–501. doi:10.1038/nbt.1630.
- Medvedev, Paul, Monica Stanciu, and Michael Brudno. 2009. "Computational Methods for Discovering Structural Variation with next-Generation Sequencing." *Nat Meth* 6 (11s). Nature Publishing Group: S13–20. <http://dx.doi.org/10.1038/nmeth.1374>.
- Menasce, L P, G R White, C J Harrison, and J M Boyle. 1993. "Localization of the Estrogen Receptor Locus (ESR) to Chromosome 6q25.1 by FISH and a Simple Post-FISH Banding Technique." *Genomics* 17 (1). UNITED STATES: 263–65. doi:10.1006/geno.1993.1320.
- Mendoza-Parra, Marco A, Martial Sankar, Mannu Walia, and Hinrich Gronemeyer. 2012. "POLYPHEMUS: R Package for Comparative Analysis of RNA Polymerase II ChIP-Seq Profiles by Non-Linear Normalization." *Nucleic Acids Research* 40 (4). England: e30. doi:10.1093/nar/gkr1205.
- Metzker, Michael L. 2010a. "Sequencing Technologies - the next Generation." *Nature Reviews. Genetics* 11 (1). England: 31–46. doi:10.1038/nrg2626.

- Meyer, Clifford A, Housheng H He, Myles Brown, and X Shirley Liu. 2011. "BINOCh: Binding Inference from Nucleosome Occupancy Changes." *Bioinformatics (Oxford, England)* 27 (13). England: 1867–68. doi:10.1093/bioinformatics/btr279.
- Mikkelsen, Tarjei S, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, et al. 2007. "Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells." *Nature* 448 (7153). England: 553–60. doi:10.1038/nature06008.
- Milne, Iain, Micha Bayer, Linda Cardle, Paul Shaw, Gordon Stephen, Frank Wright, and David Marshall. 2010. "Tablet--next Generation Sequence Assembly Visualization." *Bioinformatics (Oxford, England)* 26 (3). England: 401–2. doi:10.1093/bioinformatics/btp666.
- Montano, M M, V Muller, A Trobaugh, and B S Katzenellenbogen. 1995. "The Carboxy-Terminal F Domain of the Human Estrogen Receptor: Role in the Transcriptional Activity of the Receptor and the Effectiveness of Antiestrogens as Estrogen Antagonists." *Molecular Endocrinology (Baltimore, Md.)* 9 (7). UNITED STATES: 814–25. doi:10.1210/mend.9.7.7476965.
- Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7). United States: 621–28. doi:10.1038/nmeth.1226.
- Mosselman, S, J Polman, and R Dijkema. 1996. "ER Beta: Identification and Characterization of a Novel Human Estrogen Receptor." *FEBS Letters* 392 (1). NETHERLANDS: 49–53.
- Muino, Jose M, Kerstin Kaufmann, Roeland Chj van Ham, Gerco C Angenent, and Pawel Krajewski. 2011. "ChIP-Seq Analysis in R (CSAR): An R Package for the Statistical Detection of Protein-Bound Genomic Regions." *Plant Methods* 7. England: 11. doi:10.1186/1746-4811-7-11.
- Musgrove, Elizabeth A, and Robert L Sutherland. 2009. "Biological Determinants of Endocrine Resistance in Breast Cancer." *Nature Reviews. Cancer* 9 (9). England: 631–43. doi:10.1038/nrc2713.
- Myers, Gene. 1999. "A Fast Bit-Vector Algorithm for Approximate String Matching Based on Dynamic Programming." *J. ACM* 46 (3). New York, NY, USA: ACM: 395–415. doi:10.1145/316542.316550.
- Navakauskienė, Ruta, Grazina Treigyte, Arunas Gineitis, and Karl-Eric Magnusson. 2004. "Identification of Apoptotic Tyrosine-Phosphorylated Proteins after Etoposide or Retinoic Acid Treatment." *Proteomics* 4 (4). Germany: 1029–41. doi:10.1002/pmic.200300671.
- Negre, Nicolas, Christopher D Brown, Lijia Ma, Christopher Aaron Bristow, Steven W Miller, Ulrich Wagner, Pouya Kheradpour, et al. 2011. "A Cis-Regulatory Map of the Drosophila Genome." *Nature* 471 (7339). England: 527–31. doi:10.1038/nature09990.
- Nichols, M, J M Rientjes, and A F Stewart. 1998. "Different Positioning of the Ligand-Binding Domain Helix 12 and the F Domain of the Estrogen Receptor Accounts for Functional Differences between Agonists and Antagonists." *The EMBO Journal*. doi:10.1093/emboj/17.3.765.
- Nicholson, Robert I, and Stephen R Johnston. 2005. "Endocrine Therapy--Current Benefits and Limitations." *Breast Cancer Research and Treatment* 93 Suppl 1. Netherlands: S3–10. doi:10.1007/s10549-005-9036-4.
- Nicol, John W, Gregg A Helt, Steven G Blanchard, Archana Raja, and Ann E Loraine. 2009. "The Integrated Genome Browser: Free Software for Distribution and Exploration of Genome-Scale Datasets." *Bioinformatics*. doi:10.1093/bioinformatics/btp472.

- Ning, Z, A J Cox, and J C Mullikin. 2001. "SSAHA: A Fast Search Method for Large DNA Databases." *Genome Research* 11 (10). United States: 1725–29. doi:10.1101/gr.194201.
- Nix, David A, Samir J Courdy, and Kenneth M Boucher. 2008. "Empirical Methods for Controlling False Positives and Estimating Confidence in ChIP-Seq Peaks." *BMC Bioinformatics* 9. England: 523. doi:10.1186/1471-2105-9-523.
- Noe, Laurent, Marta Girdea, and Gregory Kucherov. 2010. "Designing Efficient Spaced Seeds for SOLiD Read Mapping." *Advances in Bioinformatics*. Egypt. doi:10.1155/2010/708501.
- Noonan, James P, and Andrew S McCallion. 2010. "Genomics of Long-Range Regulatory Elements." *Annual Review of Genomics and Human Genetics* 11. United States: 1–23. doi:10.1146/annurev-genom-082509-141651.
- Nutiu, Razvan, Robin C Friedman, Shujun Luo, Irina Khrebtukova, David Silva, Robin Li, Lu Zhang, Gary P Schroth, and Christopher B Burge. 2011. "Direct Measurement of DNA Affinity Landscapes on a High-Throughput Sequencing Instrument." *Nature Biotechnology* 29 (7). United States: 659–64. doi:10.1038/nbt.1882.
- O'Neill, Laura P, Matthew D VerMilyea, and Bryan M Turner. 2006. "Epigenetic Characterization of the Early Embryo with a Chromatin Immunoprecipitation Protocol Applicable to Small Cell Populations." *Nature Genetics* 38 (7). United States: 835–41. doi:10.1038/ng1820.
- Odom, Duncan T, Robin D Dowell, Elizabeth S Jacobsen, William Gordon, Timothy W Danford, Kenzie D MacIsaac, P Alexander Rolfe, Caitlin M Conboy, David K Gifford, and Ernest Fraenkel. 2007. "Tissue-Specific Transcriptional Regulation Has Diverged Significantly between Human and Mouse." *Nature Genetics* 39 (6). United States: 730–32. doi:10.1038/ng2047.
- Osborne, C K. 1998. "Steroid Hormone Receptors in Breast Cancer Management." *Breast Cancer Research and Treatment* 51 (3). NETHERLANDS: 227–38.
- Osborne, C K, A Wakeling, and R I Nicholson. 2004. "Fulvestrant: An Oestrogen Receptor Antagonist with a Novel Mechanism of Action." *British Journal of Cancer*. doi:10.1038/sj.bjc.6601629.
- Osborne, C Kent, and Rachel Schiff. 2011. "Mechanisms of Endocrine Resistance in Breast Cancer." *Annual Review of Medicine* 62. United States: 233–47. doi:10.1146/annurev-med-070909-182917.
- Owolabi, O, and D R McGregor. 1988. "Fast Approximate String Matching." *Software: Practice and Experience* 18 (4). John Wiley & Sons, Ltd.: 387–93. doi:10.1002/spe.4380180407.
- Pabinger, Stephan, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R Speicher, Johannes Zschocke, and Zlatko Trajanoski. 2013. "A Survey of Tools for Variant Analysis of next-Generation Genome Sequencing Data." *Briefings in Bioinformatics*, January. doi:10.1093/bib/bbs086.
- Park, Peter J. 2009. "ChIP-Seq: Advantages and Challenges of a Maturing Technology." *Nature Reviews. Genetics* 10 (10). England: 669–80. doi:10.1038/nrg2641.
- Peng, Shouyong, Artyom A Alekseyenko, Erica Larschan, Mitzi I Kuroda, and Peter J Park. 2007. "Normalization and Experimental Design for ChIP-Chip Data." *BMC Bioinformatics* 8. England: 219. doi:10.1186/1471-2105-8-219.
- Pepke, Shirley, Barbara Wold, and Ali Mortazavi. 2009. "Computation for ChIP-Seq and RNA-Seq Studies." *Nature Methods* 6 (11 Suppl). United States: S22–32. doi:10.1038/nmeth.1371.
- Pickrell, Joseph K, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. 2011. "False Positive Peaks in ChIP-Seq and Other Sequencing-Based Functional Assays Caused by Unannotated High Copy Number Regions." *Bioinformatics (Oxford, England)* 27 (15). England: 2144–46. doi:10.1093/bioinformatics/btr354.

- Planet, Evarist, Camille Stephan-Otto Attolini, Oscar Reina, Oscar Flores, and David Rossell. 2012. "htSeqTools: High-Throughput Sequencing Quality Control, Processing and Visualization in R." *Bioinformatics (Oxford, England)* 28 (4). England: 589–90. doi:10.1093/bioinformatics/btr700.
- Podicheti, Ram, and Qunfeng Dong. 2011. "Administering GBrowse Sites with WebGBrowse." *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxeavanis ... [et Al.]* Chapter 9 (March). United States: Unit 9.14. doi:10.1002/0471250953.bi0914s33.
- Qin, Zhaohui S, Jianjun Yu, Jincheng Shen, Christopher A Maher, Ming Hu, Shanker Kalyana-Sundaram, Jindan Yu, and Arul M Chinnaiyan. 2010. "HPeak: An HMM-Based Algorithm for Defining Read-Enriched Regions in ChIP-Seq Data." *BMC Bioinformatics* 11. England: 369. doi:10.1186/1471-2105-11-369.
- Quail, Michael A, Iwanka Kozarewa, Frances Smith, Aylwyn Scally, Philip J Stephens, Richard Durbin, Harold Swerdlow, and Daniel J Turner. 2008. "A Large Genome Center's Improvements to the Illumina Sequencing System." *Nature Methods* 5 (12). United States: 1005–10. doi:10.1038/nmeth.1270.
- Quinlan, Aaron R, and Ira M Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics (Oxford, England)* 26 (6). England: 841–42. doi:10.1093/bioinformatics/btq033.
- Raney, Brian J, Melissa S Cline, Kate R Rosenbloom, Timothy R Dreszer, Katrina Learned, Galt P Barber, Laurence R Meyer, et al. 2011. "ENCODE Whole-Genome Data in the UCSC Genome Browser (2011 Update)." *Nucleic Acids Research*. doi:10.1093/nar/gkq1017.
- Rashid, Naim U, Paul G Giresi, Joseph G Ibrahim, Wei Sun, and Jason D Lieb. 2011. "ZINBA Integrates Local Covariates with DNA-Seq Data to Identify Broad and Narrow Regions of Enrichment, Even within Amplified Genomic Regions." *Genome Biology* 12 (7). England: R67. doi:10.1186/gb-2011-12-7-r67.
- Rasmussen, Kim R, Jens Stoye, and Eugene W Myers. 2006. "Efficient Q-Gram Filters for Finding All Epsilon-Matches over a given Length." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 13 (2). United States: 296–308. doi:10.1089/cmb.2006.13.296.
- Ravasi, Timothy, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, et al. 2010. "An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man." *Cell* 140 (5). United States: 744–52. doi:10.1016/j.cell.2010.01.044.
- Ren, B, F Robert, J J Wyrick, O Aparicio, E G Jennings, I Simon, J Zeitlinger, et al. 2000. "Genome-Wide Location and Function of DNA Binding Proteins." *Science (New York, N.Y.)* 290 (5500). UNITED STATES: 2306–9. doi:10.1126/science.290.5500.2306.
- Rhodes, Daniel R, Jianjun Yu, K Shanker, Nandan Deshpande, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pandey, and Arul M Chinnaiyan. 2004. "ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform." *Neoplasia (New York, N.Y.)*.
- Riggins, Rebecca B, Randy S Schrecengost, Michael S Guerrero, and Amy H Bouton. 2007. "Pathways to Tamoxifen Resistance." *Cancer Letters* 256 (1). Ireland: 1–24. doi:10.1016/j.canlet.2007.03.016.
- Ring, Alistair, and Mitch Dowsett. 2004. "Mechanisms of Tamoxifen Resistance." *Endocrine-Related Cancer* 11 (4). England: 643–58. doi:10.1677/erc.1.00776.

- Ringner, Markus, Erik Fredlund, Jari Hakkinen, Ake Borg, and Johan Staaf. 2011. "GOBO: Gene Expression-Based Outcome for Breast Cancer Online." *PLoS One* 6 (3). United States: e17911. doi:10.1371/journal.pone.0017911.
- Rizk, Guillaume, and Dominique Lavenier. 2010. "GASSST: Global Alignment Short Sequence Search Tool." *Bioinformatics* 26 (20): 2534–40. doi:10.1093/bioinformatics/btq485.
- Robasky, Kimberly, and Martha L Bulyk. 2011. "UniPROBE, Update 2011: Expanded Content and Search Tools in the Online Database of Protein-Binding Microarray Data on Protein-DNA Interactions." *Nucleic Acids Research* 39 (Database issue). England: D124–28. doi:10.1093/nar/gkq992.
- Robertson, Gordon, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, et al. 2007. "Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing." *Nature Methods* 4 (8). United States: 651–57. doi:10.1038/nmeth1068.
- Robinson, James T, Helga Thorvaldsdottir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. "Integrative Genomics Viewer." *Nature Biotechnology*. United States. doi:10.1038/nbt.1754.
- Roider, Helge G, Aditi Kanhere, Thomas Manke, and Martin Vingron. 2007. "Predicting Transcription Factor Affinities to DNA from a Biophysical Model." *Bioinformatics (Oxford, England)* 23 (2). England: 134–41. doi:10.1093/bioinformatics/btl565.
- Roider, Helge G, Thomas Manke, Sean O’Keeffe, Martin Vingron, and Stefan A Haas. 2009. "PASTAA: Identifying Transcription Factors Associated with Sets of Co-Regulated Genes." *Bioinformatics*. doi:10.1093/bioinformatics/btn627.
- Rosenbloom, Kate R, Timothy R Dreszer, Jeffrey C Long, Venkat S Malladi, Cricket A Sloan, Brian J Raney, Melissa S Cline, et al. 2012. "ENCODE Whole-Genome Data in the UCSC Genome Browser: Update 2012." *Nucleic Acids Research* 40 (Database issue). England: D912–17. doi:10.1093/nar/gkr1012.
- Ross-Innes, Caryn S, Rory Stark, Andrew E Teschendorff, Kelly A Holmes, H Raza Ali, Mark J Dunning, Gordon D Brown, et al. 2012. "Differential Oestrogen Receptor Binding Is Associated with Clinical Outcome in Breast Cancer." *Nature* 481 (7381). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 389–93. <http://dx.doi.org/10.1038/nature10730>.
- Rozowsky, Joel, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. 2009. "PeakSeq Enables Systematic Scoring of ChIP-Seq Experiments Relative to Controls." *Nat Biotech* 27 (1). Nature Publishing Group: 66–75. <http://dx.doi.org/10.1038/nbt.1518>.
- Rumble, Stephen M, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. 2009. "SHRiMP: Accurate Mapping of Short Color-Space Reads." *PLoS Computational Biology* 5 (5). United States: e1000386. doi:10.1371/journal.pcbi.1000386.
- Russo, J, X Ao, C Grill, and I H Russo. 1999. "Pattern of Distribution of Cells Positive for Estrogen Receptor Alpha and Progesterone Receptor in Relation to Proliferating Cells in the Mammary Gland." *Breast Cancer Research and Treatment* 53 (3). NETHERLANDS: 217–27.
- Sabo, Peter J, Richard Humbert, Michael Hawrylycz, James C Wallace, Michael O Dorschner, Michael McArthur, and John A Stamatoyannopoulos. 2004. "Genome-Wide Identification of DNaseI Hypersensitive Sites Using Active Chromatin Sequence Libraries." *Proceedings of the National Academy of Sciences of the United States of America* 101 (13). United States: 4537–42. doi:10.1073/pnas.0400678101.

- Safe, Stephen, and Maen Abdelrahim. 2005. "Sp Transcription Factor Family and Its Role in Cancer." *European Journal of Cancer (Oxford, England : 1990)* 41 (16). England: 2438–48. doi:10.1016/j.ejca.2005.08.006.
- Salmon-Divon, Mali, Heidi Dvinge, Kairi Tammoja, and Paul Bertone. 2010. "PeakAnalyzer: Genome-Wide Annotation of Chromatin Binding and Modification Loci." *BMC Bioinformatics*. doi:10.1186/1471-2105-11-415.
- Sandelin, Albin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. 2004. "JASPAR: An Open-Access Database for Eukaryotic Transcription Factor Binding Profiles." *Nucleic Acids Research*. Oxford, UK. doi:10.1093/nar/gkh012.
- Sanger, F, and A R Coulson. 1975. "A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase." *Journal of Molecular Biology* 94 (3): 441–48. citeulike-article-id:2308945.
- Santen, Richard J, Norman F Boyd, Rowan T Chlebowski, Steven Cummings, Jack Cuzick, Mitch Dowsett, Douglas Easton, et al. 2007. "Critical Assessment of New Risk Factors for Breast Cancer: Considerations for Development of an Improved Risk Prediction Model." *Endocrine-Related Cancer* 14 (2). England: 169–87. doi:10.1677/ERC-06-0045.
- Schmidt, Dominic, Michael D Wilson, Benoit Ballester, Petra C Schwalie, Gordon D Brown, Aileen Marshall, Claudia Kutter, et al. 2010. "Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding." *Science (New York, N.Y.)* 328 (5981). United States: 1036–40. doi:10.1126/science.1186176.
- Schmieder, Robert, and Robert Edwards. 2011. "Quality Control and Preprocessing of Metagenomic Datasets." *Bioinformatics (Oxford, England)* 27 (6). England: 863–64. doi:10.1093/bioinformatics/btr026.
- Schmieder, Robert, Yan Wei Lim, Forest Rohwer, and Robert Edwards. 2010. "TagCleaner: Identification and Removal of Tag Sequences from Genomic and Metagenomic Datasets." *BMC Bioinformatics*. doi:10.1186/1471-2105-11-341.
- Schneider, T D, G D Stormo, L Gold, and A Ehrenfeucht. 1986. "Information Content of Binding Sites on Nucleotide Sequences." *Journal of Molecular Biology* 188 (3). ENGLAND: 415–31.
- Schultz, Jennifer R, Larry N Petz, and Ann M Nardulli. 2005. "Cell- and Ligand-Specific Regulation of Promoters Containing Activator Protein-1 and Sp1 Sites by Estrogen Receptors Alpha and Beta." *The Journal of Biological Chemistry* 280 (1). United States: 347–54. doi:10.1074/jbc.M407879200.
- Schuster, Stephan C. 2008. "Next-Generation Sequencing Transforms Today's Biology." *Nature Methods* 5 (1). United States: 16–18. doi:10.1038/nmeth1156.
- Schuster, Stephan C, Webb Miller, Aakrosh Ratan, Lynn P Tomsho, Belinda Giardine, Lindsay R Kasson, Robert S Harris, et al. 2010. "Complete Khoisan and Bantu Genomes from Southern Africa." *Nature* 463 (7283). England: 943–47. doi:10.1038/nature08795.
- Schwappacher, Raphaela, Hema Rangaswami, Jacqueline Su-Yuo, Aaron Hassad, Ryan Spitler, and Darren E Casteel. 2013. "cGMP-Dependent Protein Kinase I β Regulates Breast Cancer Cell Migration and Invasion via Interaction with the Actin/myosin-Associated Protein Caldesmon." *Journal of Cell Science*. Bidder Building, 140 Cowley Road, Cambridge, CB4 0DL, UK. doi:10.1242/jcs.118190.

- Schwartz, Janice A, Li Zhong, Sarah Deighton-Collins, Changqing Zhao, and Debra F Skafar. 2002. "Mutations Targeted to a Predicted Helix in the Extreme Carboxyl-Terminal Region of the Human Estrogen Receptor-Alpha Alter Its Response to Estradiol and 4-Hydroxytamoxifen." *The Journal of Biological Chemistry* 277 (15). United States: 13202–9. doi:10.1074/jbc.M112215200.
- Shao, Zhen, Yijing Zhang, Guo-Cheng Yuan, Stuart H Orkin, and David J Waxman. 2012. "MAnorm: A Robust Model for Quantitative Comparison of ChIP-Seq Data Sets." *Genome Biology* 13 (3). England: R16. doi:10.1186/gb-2012-13-3-r16.
- Sharov, Alexei A, and Minoru S H Ko. 2009. "Exhaustive Search for Over-Represented DNA Sequence Motifs with CisFinder." *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*. doi:10.1093/dnares/dsp014.
- Shefer-Vaida, M, A Sherman, T Ashkenazi, K Robzyk, and Y Kassir. 1995. "Positive and Negative Feedback Loops Affect the Transcription of IME1, a Positive Regulator of Meiosis in *Saccharomyces Cerevisiae*." *Developmental Genetics* 16 (3). UNITED STATES: 219–28. doi:10.1002/dvg.1020160302.
- Shin, Hyunjin, Tao Liu, Arjun K Manrai, and X Shirley Liu. 2009. "CEAS: Cis-Regulatory Element Annotation System." *Bioinformatics (Oxford, England)* 25 (19). England: 2605–6. doi:10.1093/bioinformatics/btp479.
- Shin, Hyunjin, Tao Liu, Xikun Duan, Yong Zhang, and X Shirley Liu. 2013. "Computational Methodology for ChIP-Seq Analysis." *Quantitative Biology* 1 (1): 54–70. doi:10.1007/s40484-013-0006-2.
- Smagulova, Fatima. 2011. "Genome-Wide Analysis Reveals Novel Molecular Features of Mouse Recombination Hotspots." doi:10.1038/nature09869.
- Solomon, M J, P L Larsen, and A Varshavsky. 1988. "Mapping Protein-DNA Interactions in Vivo with Formaldehyde: Evidence That Histone H4 Is Retained on a Highly Transcribed Gene." *Cell* 53 (6). UNITED STATES: 937–47.
- Spyrou, Christiana, Rory Stark, Andy G Lynch, and Simon Tavaré. 2009. "BayesPeak: Bayesian Analysis of ChIP-Seq Data." *BMC Bioinformatics* 10. England: 299. doi:10.1186/1471-2105-10-299.
- Storey, John D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3). Blackwell Publishers: 479–98. doi:10.1111/1467-9868.00346.
- Storey, John D, and Robert Tibshirani. 2003. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences of the United States of America* 100 (16). United States: 9440–45. doi:10.1073/pnas.1530509100.
- Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43). United States: 15545–50. doi:10.1073/pnas.0506580102.
- Suzuki, Mitsuko, Naoto Ueno, and Atsushi Kuroiwa. 2003. "Hox Proteins Functionally Cooperate with the GC Box-Binding Protein System through Distinct Domains." *The Journal of Biological Chemistry* 278 (32). United States: 30148–56. doi:10.1074/jbc.M303932200.

- Suzuki, Shingo, Naoaki Ono, Chikara Furusawa, Bei-Wen Ying, and Tetsuya Yomo. 2011. "Comparison of Sequence Reads Obtained from Three next-Generation Sequencing Platforms." *PloS One* 6 (5). United States: e19534. doi:10.1371/journal.pone.0019534.
- Szalkowski, Adam M, and Christoph D Schmid. 2011. "Rapid Innovation in ChIP-Seq Peak-Calling Algorithms Is Outdistancing Benchmarking Efforts." *Briefings in Bioinformatics* 12 (6). England: 626–33. doi:10.1093/bib/bbq068.
- Tamrazi, Anobel, Kathryn E Carlson, Jonathan R Daniels, Kyle M Hurth, and John A Katzenellenbogen. 2002. "Estrogen Receptor Dimerization: Ligand Binding Regulates Dimer Affinity and Dimer Dissociation Rate." *Molecular Endocrinology (Baltimore, Md.)* 16 (12). United States: 2706–19. doi:10.1210/me.2002-0250.
- Tan, Juan, Chen-Yang Yu, Zhen-Hua Wang, Hao-Yan Chen, Jian Guan, Ying-Xuan Chen, and Jing-Yuan Fang. 2015. "Genetic Variants in the Inositol Phosphate Metabolism Pathway and Risk of Different Types of Cancer." *Scientific Reports* 5. England: 8473. doi:10.1038/srep08473.
- Tang, Qianzi, Yiwen Chen, Clifford Meyer, Tim Geistlinger, Mathieu Lupien, Qian Wang, Tao Liu, Yong Zhang, Myles Brown, and Xiaole Shirley Liu. 2011. "A Comprehensive View of Nuclear Receptor Cancer Cistromes." *Cancer Research* 71 (22). United States: 6940–47. doi:10.1158/0008-5472.CAN-11-2091.
- Taslim, Cenny, Tim Huang, and Shili Lin. 2011. "DIME: R-Package for Identifying Differential ChIP-Seq Based on an Ensemble of Mixture Models." *Bioinformatics*. doi:10.1093/bioinformatics/btr165.
- Taslim, Cenny, Jiejun Wu, Pearly Yan, Greg Singer, Jeffrey Parvin, Tim Huang, Shili Lin, and Kun Huang. 2009. "Comparative Study on ChIP-Seq Data: Normalization and Binding Pattern Characterization." *Bioinformatics*. doi:10.1093/bioinformatics/btp384.
- Tokunaga, Eriko, Yasue Kimura, Kojiro Mashino, Eiji Oki, Akemi Kataoka, Shinji Ohno, Masaru Morita, Yoshihiro Kakeji, Hideo Baba, and Yoshihiko Maehara. 2006. "Activation of PI3K/Akt Signaling and Hormone Resistance in Breast Cancer." *Breast Cancer (Tokyo, Japan)* 13 (2). Japan: 137–44.
- Theocharis, Achilleas D, Spyros S Skandalis, Thomas Neill, Hinke A B Mulhaupt, Mario Hubo, Helena Frey, Sandeep Gopal, et al. 2015. "Insights into the Key Roles of Proteoglycans in Breast Cancer Biology and Translational Medicine." *Biochimica et Biophysica Acta* 1855 (2): 276–300. doi:10.1016/j.bbcan.2015.03.006.
- Thomas, Paul D, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. 2003. "PANTHER: A Library of Protein Families and Subfamilies Indexed by Function." *Genome Research* 13 (9). United States: 2129–41. doi:10.1101/gr.772403.
- Thomas-Chollier, Morgane, Carl Herrmann, Matthieu Defrance, Olivier Sand, Denis Thieffry, and Jacques van Helden. 2012. "RSAT Peak-Motifs: Motif Analysis in Full-Size ChIP-Seq Datasets." *Nucleic Acids Research* 40 (4). England: e31. doi:10.1093/nar/gkr1104.
- Thompson, William, Eric C Rouchka, and Charles E Lawrence. 2003. "Gibbs Recursive Sampler: Finding Transcription Factor Binding Sites." *Nucleic Acids Research* 31 (13). England: 3580–85.

- Tora, Laszio, John White, Christel Brou, Diane Tasset, Nicholas Webster, Elisabeth Scheer, and Pierre Chambon. 1989. "The Human Estrogen Receptor Has Two Independent Nonacidic Transcriptional Activation Functions." *Cell* 59 (3): 477–87. doi:http://dx.doi.org/10.1016/0092-8674(89)90031-7.
- Toth, J, and M D Biggin. 2000. "The Specificity of Protein-DNA Crosslinking by Formaldehyde: In Vitro and in Drosophila Embryos." *Nucleic Acids Research* 28 (2). ENGLAND: e4.
- Trapnell, Cole, and Steven L Salzberg. 2009. "How to Map Billions of Short Reads onto Genomes." *Nature Biotechnology*. doi:10.1038/nbt0509-455.
- Tuerk, C, and L Gold. 1990. "Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase." *Science (New York, N.Y.)* 249 (4968). UNITED STATES: 505–10.
- Tuteja, Geetu, Peter White, Jonathan Schug, and Klaus H Kaestner. 2009. "Extracting Transcription Factor Targets from ChIP-Seq Data." *Nucleic Acids Research* 37 (17). England: e113. doi:10.1093/nar/gkp536.
- Uzman, A. 2001. "Molecular Cell Biology (4th Edition) - Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore and James Darnell; Freeman & Co., New York, NY, 2000, 1084 Pp., List Price \$102.25, ISBN 0-7167-3136-3." *Biochemistry and Molecular Biology Education*. Elsevier, 126–28. doi:doi: 10.1016/s1470-8175(01)00023-6.
- Valouev, Anton, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. 2008. "Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-Seq Data." *Nature Methods* 5 (9). United States: 829–34. doi:10.1038/nmeth.1246.
- Van Helden, Jacques. 2003. "Regulatory Sequence Analysis Tools." *Nucleic Acids Research* 31 (13). England: 3593–96.
- Verzi, Michael P, Hyunjin Shin, H Hansen He, Rita Sulahian, Clifford A Meyer, Robert K Montgomery, James C Fleet, Myles Brown, X Shirley Liu, and Ramesh A Shivdasani. 2010. "Differentiation-Specific Histone Modifications Reveal Dynamic Chromatin Interactions and Partners for the Intestinal Transcription Factor CDX2." *Developmental Cell* 19 (5). United States: 713–26. doi:10.1016/j.devcel.2010.10.006.
- Via, Marc, Christopher Gignoux, and Esteban González Burchard. 2010. "The 1000 Genomes Project: New Opportunities for Research and Social Challenges." *Genome Medicine*. doi:10.1186/gm124.
- Von Bubnoff, Andreas. 2015. "Next-Generation Sequencing: The Race Is On." *Cell* 132 (5). Elsevier: 721–23. doi:10.1016/j.cell.2008.02.028.
- Wakeling, A E, M Dukes, and J Bowler. 1991. "A Potent Specific Pure Antiestrogen with Clinical Potential." *Cancer Research* 51 (15). UNITED STATES: 3867–73.
- Walter, P, S Green, G Greene, A Krust, J M Bornert, J M Jeltsch, A Staub, E Jensen, G Scrase, and M Waterfield. 1985. "Cloning of the Human Estrogen Receptor cDNA." *Proceedings of the National Academy of Sciences of the United States of America*.
- Wang, Congmao, Jie Xu, Dasheng Zhang, Zoe A Wilson, and Dabing Zhang. 2010. "An Effective Approach for Identification of in Vivo Protein-DNA Binding Sites from Paired-End ChIP-Seq Data." *BMC Bioinformatics* 11. England: 81. doi:10.1186/1471-2105-11-81.

- Wang, Ting, Jue Zeng, Craig B Lowe, Robert G Sellers, Sofie R Salama, Min Yang, Shawn M Burgess, Rainer K Brachmann, and David Haussler. 2007. "Species-Specific Endogenous Retroviruses Shape the Transcriptional Network of the Human Tumor Suppressor Protein p53." *Proceedings of the National Academy of Sciences of the United States of America* 104 (47). United States: 18613–18. doi:10.1073/pnas.0703637104.
- Wang, Zhibin, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, et al. 2008. "Combinatorial Patterns of Histone Acetylations and Methylations in the Human Genome." *Nature Genetics* 40 (7). United States: 897–903. doi:10.1038/ng.154.
- Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews. Genetics* 10 (1). England: 57–63. doi:10.1038/nrg2484.
- Weese, David, Anne-Katrin Emde, Tobias Rausch, Andreas Doring, and Knut Reinert. 2009. "RazerS-Fast Read Mapping with Sensitivity Control." *Genome Research* 19 (9). United States: 1646–54. doi:10.1101/gr.088823.108.
- Wei, Chia-Lin, Qiang Wu, Vinsensius B Vega, Kuo Ping Chiu, Patrick Ng, Tao Zhang, Atif Shahab, et al. 2006. "A Global Map of p53 Transcription-Factor Binding Sites in the Human Genome." *Cell* 124 (1). United States: 207–19. doi:10.1016/j.cell.2005.10.043.
- White, R, M Sjöberg, E Kalkhoven, and M G Parker. 1997. "Ligand-Independent Activation of the Oestrogen Receptor by Mutation of a Conserved Tyrosine." *The EMBO Journal* 16 (6). ENGLAND: 1427–35. doi:10.1093/emboj/16.6.1427.
- Whittington, Tom, Martin C Frith, James Johnson, and Timothy L Bailey. 2011. "Inferring Transcription Factor Complexes from ChIP-Seq Data." *Nucleic Acids Research* 39 (15). England: e98. doi:10.1093/nar/gkr341.
- Wilbanks, Elizabeth G, and Marc T Facciotti. 2010. "Evaluation of Algorithm Performance in ChIP-Seq Peak Detection." *PloS One* 5 (7). United States: e11471. doi:10.1371/journal.pone.0011471.
- Williams, Kristine, Jesper Christensen, Marianne Terndrup Pedersen, Jens V Johansen, Paul A C Cloos, Juri Rappsilber, and Kristian Helin. 2011. "TET1 and Hydroxymethylcytosine in Transcription and DNA Methylation Fidelity." *Nature* 473 (7347). England: 343–48. doi:10.1038/nature10066.
- Wiseman, H, and R Duffy. 2001. "New Advances in the Understanding of the Role of Steroids and Steroid Receptors in Disease." *Biochemical Society Transactions* 29 (Pt 2). England: 205–9.
- Wold, Barbara, and Richard M Myers. 2008. "Sequence Census Methods for Functional Genomics." *Nature Methods* 5 (1). United States: 19–21. doi:10.1038/nmeth1157.
- Wu, Angela R, Joseph B Hiatt, Rong Lu, Joanne L Attema, Neethan A Lobo, Irving L Weissman, Michael F Clarke, and Stephen R Quake. 2009. "Automated Microfluidic Chromatin Immunoprecipitation from 2,000 Cells." *Lab on a Chip* 9 (10). England: 1365–70. doi:10.1039/b819648f.
- Wu, Hao, and Hongkai Ji. 2014a. "PolyaPeak: Detecting Transcription Factor Binding Sites from ChIP-Seq Using Peak Shape Information." *PloS One* 9 (3). United States: e89694. doi:10.1371/journal.pone.0089694.
- Wu, Thomas D, and Serban Nacu. 2010. "Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads." *Bioinformatics* 26 (7): 873–81. doi:10.1093/bioinformatics/btq057.

- Xu, Han, Lusy Handoko, Xueliang Wei, Chaopeng Ye, Jianpeng Sheng, Chia-Lin Wei, Feng Lin, and Wing-Kin Sung. 2010. "A Signal-Noise Model for Significance Analysis of ChIP-Seq with Negative Control." *Bioinformatics (Oxford, England)* 26 (9). England: 1199–1204. doi:10.1093/bioinformatics/btq128.
- Xu, Han, Chia-Lin Wei, Feng Lin, and Wing-Kin Sung. 2008. "An HMM Approach to Genome-Wide Identification of Differential Histone Modification Sites from ChIP-Seq Data." *Bioinformatics* 24 (20): 2344–49. <http://bioinformatics.oxfordjournals.org/content/24/20/2344.abstract>.
- Xu, Shiliyang, Sean Grullon, Kai Ge, and Weiqun Peng. 2014. "Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells." *Methods in Molecular Biology (Clifton, N.J.)* 1150. United States: 97–111. doi:10.1007/978-1-4939-0512-6_5.
- Yamaguchi, Hideki, and John Condeelis. 2007. "Regulation of the Actin Cytoskeleton in Cancer Cell Migration and Invasion." *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1773 (5): 642–52. doi:<http://dx.doi.org/10.1016/j.bbamcr.2006.07.001>.
- Yamauchi, Masashi, Shinji Kawai, Takahiro Kato, Takashi Ooshima, and Atsuo Amano. 2008. "Odd-Skipped Related 1 Gene Expression Is Regulated by Runx2 and Ikzf1 Transcription Factors." *Gene* 426 (1-2). Netherlands: 81–90. doi:10.1016/j.gene.2008.08.015.
- Yarden, Y, and M X Sliwkowski. 2001. "Untangling the ErbB Signalling Network." *Nature Reviews. Molecular Cell Biology* 2 (2). England: 127–37. doi:10.1038/35052073.
- Ylikomi, T, M T Bocquel, M Berry, H Gronemeyer, and P Chambon. 1992. "Cooperation of Proto-Signals for Nuclear Accumulation of Estrogen and Progesterone Receptors." *The EMBO Journal*.
- Young, Daniel W, Mohammad Q Hassan, Xiao-Qing Yang, Mario Galindo, Amjad Javed, Sayyed K Zaidi, Paul Furcinitti, et al. 2007. "Mitotic Retention of Gene Expression Patterns by the Cell Fate-Determining Transcription Factor Runx2." *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.0611419104.
- Yu, Ming, Laura Riva, Huafeng Xie, Yocheved Schindler, Tyler B Moran, Yong Cheng, Duonan Yu, et al. 2009. "Insights into GATA-1-Mediated Gene Activation versus Repression via Genome-Wide Chromatin Occupancy Analysis." *Molecular Cell* 36 (4). United States: 682–95. doi:10.1016/j.molcel.2009.11.002.
- Yu, Xiaoqing, Kishore Guda, Joseph Willis, Martina Veigl, Zhenghe Wang, Sanford Markowitz, Mark D Adams, and Shuying Sun. 2012a. "How Do Alignment Programs Perform on Sequencing Data with Varying Qualities and from Repetitive Regions?" *BioData Mining* 5 (1). England: 6. doi:10.1186/1756-0381-5-6.
- Yuh, C H, H Bolouri, and E H Davidson. 1998. "Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene." *Science (New York, N.Y.)* 279 (5358). UNITED STATES: 1896–1902.
- Zang, Chongzhi, Dustin E Schones, Chen Zeng, Kairong Cui, Keji Zhao, and Weiqun Peng. 2009. "A Clustering Approach for Identification of Enriched Domains from Histone Modification ChIP-Seq Data." *Bioinformatics (Oxford, England)* 25 (15). England: 1952–58. doi:10.1093/bioinformatics/btp340.
- Zhang, Jun, Rod Chiodini, Ahmed Badr, and Genfa Zhang. 2011. "The Impact of next-Generation Sequencing on Genomics." *Journal of Genetics and Genomics = Yi Chuan Xue Bao* 38 (3). China: 95–109. doi:10.1016/j.jgg.2011.02.003.

- Zhang, Xuekui, Gordon Robertson, Martin Krzywinski, Kaida Ning, Arnaud Droit, Steven Jones, and Raphael Gottardo. 2011. "PICS: Probabilistic Inference for ChIP-Seq." *Biometrics* 67 (1). Blackwell Publishing Inc: 151–63. doi:10.1111/j.1541-0420.2010.01441.x.
- Zhang, Yong, Tao Liu, Clifford A Meyer, Jerome Eeckhoutte, David S Johnson, Bradley E Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9). England: R137. doi:10.1186/gb-2008-9-9-r137.
- Zhang, Yong, Hyunjin Shin, Jun S Song, Ying Lei, and X Shirley Liu. 2008. "Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq." *BMC Genomics* 9. England: 537. doi:10.1186/1471-2164-9-537.
- Zhang, Zhizhuo, Cheng Wei Chang, Wan Ling Goh, Wing-Kin Sung, and Edwin Cheung. 2011. "CENTDIST: Discovery of Co-Associated Factors by Motif Distribution." *Nucleic Acids Research* 39 (Web Server issue). England: W391–99. doi:10.1093/nar/gkr387.
- Zheng, Yiyu, Xiaoman Li, and Haiyan Hu. 2015. "PreDREM: A Database of Predicted DNA Regulatory Motifs from 349 Human Cell and Tissue Samples." *Database: The Journal of Biological Databases and Curation* 2015 (February). Oxford University Press: bav007. doi:10.1093/database/bav007.
- Zhou, Y H, D Hewett-Emmett, J P Ward, and W H Li. 1997. "Unexpected Conservation of the X-Linked Color Vision Gene in Nocturnal Prosimians: Evidence from Two Bush Babies." *Journal of Molecular Evolution* 45 (6). UNITED STATES: 610–18.